A Spatial Display for Ground-Penetrating Radar Change Detection

by

Paul W. Quimby

S.B., Massachusetts Institute of Technology (2012)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

Author ....
Department of
Electrical Engineering and Computer Science
September 1, 2013

Certified by..
Mary L. Cummings
Visiting Professor of Aeronautics and Astronautics
Thesis Supervisor

Accepted by ....
Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

A Spatial Display for Ground-Penetrating Radar Change Detection

by

Paul W. Quimby

Submitted to the Department of
Electrical Engineering and Computer Science
on September 1, 2013, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Ground-Penetrating Radar (GPR) enables the exploration and mapping of subterranean volumes for applications such as construction, humanitarian demining, archeology, and environmental science. In each of these applications, special signal processing pipelines have been developed to reduce noise and reject clutter for optimal object detection and tracking. Change Detection (CD) is one approach to solving these signal detection challenges by leveraging the concept that changes are more relevant than absolute measurements. This research focuses on the Gopher vehicle-mounted CD GPR system.

Regardless of the application, GPR data must be interpreted by some intelligence, whether human or artificial. Traditional GPR interfaces present the raw GPR data to an operator in cross sections organized by time and depth. The intent of these displays is to allow a human operator to formulate a mental model and plan of action. After a human factors evaluation, this presentation was identified as suboptimal, and a new display was designed to present GPR data. The new display organizes data in a spatial manner and presents the information to the operator on a map. The display was tested in a human subjects experiment with thirty untrained volunteers and two expert operators measuring the signal detection properties of the display compared with a traditional temporally-organized display. The display and operator system was evaluated using signal detection theory analysis.

The new spatial display was quantitatively superior as evidenced by a 4.7% increase in correctness of the subjects' classifications and a 29% decrease in miss percentage. Qualitatively, 83% of subjects preferred the new interface and 7% had no preference. The collective intelligence implications of this system were investigated by simulating voting committees of operators. Committee performance was superior to expert operators and to top performers in several respects.

Thesis Supervisor: Mary L. Cummings
Title: Visiting Professor of Aeronautics and Astronautics

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

Acknowledgments

Research does not and should not happen in a vacuum. The people I have been fortunate enough to work with over the last several years have been influential in making me the person I am today and the investigator I attempted to be in my work.

I would like to thank Missy Cummings, my research advisor at MIT. I stumbled upon HAL my freshman year and decided to stick around. It was an excellent decision, and I have been very lucky to have had the opportunity to work for you. Your willingness to develop the skills of the people working for you, while simultaneously reminding us of our inadequacies is a powerful combination that drove me to do better. Your efforts to create a productive and diverse lab have been very successful, and I am saddened to be leaving.

Thank you, Matt Cornick, for your guidance and support throughout this process. I have enjoyed getting to work on this project immensely. Working with you and the project team has been a great chance to experience a new level of engineering expertise, and I am very glad I picked this project.

Thank you to everyone I have had the pleasure of working with in HAL. Particularly Alex, Andrew, Andrew, Dave, Erin, Jason, Kathleen, Kim, Luca, and Yves.

Sally Chapman, Jennifer Craig, and Anne Hunter, you have each saved me from gruesome fates (several times) at the hands of the Institute and academic life. Thank you also to the chaplains and deans in student support services.

I would like to thank Dr. Lee Spence, Dr. Justin Brooke, and Dr. Robert Shin for helping me find and arrange the financial support which made this research possible. Thank you all for taking the time to encourage me as a new member of the Lincoln community.

Thank you, Rebecca, for being you and being here with me at MIT.

I would like to thank my parents for their support and guidance. I am blessed to have grown up in your house and learned so much in your care.

THIS PAGE INTENTIONALLY LEFT BLANK

Nomenclature

CD Change Detection 17

DBW Drive-By-Wire 18

EMI Electromagnetic Induction 15

FA False Alarm 29

FPR False Positive Rate 29

GPR Ground-Penetrating Radar 15

hCTA hybrid Cognitive Task Analysis 36

PD Probability of Detection 29

ROC Receiver Operating Characteristic 30

TPR True Positive Rate 29

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

This chapter will present an introduction to this thesis. It will first explain the basics of Ground-Penetrating Radar (GPR) and the specific system that was studied, Gopher. Then the motivation for why better GPR systems are required will be discussed along with the challenges of GPR processing. The criteria for success in this research will be presented in summary of the human factors challenges relevant to a good interface solution. At the end of this chapter, the research goals of this work and the rest of the thesis will be outlined.

1.1 Ground-Penetrating Radar

Radar can be used to sense and model the world with greater accuracy and reliability than many other types of sensors. Since its development during WWII by Allied engineers, microwave radar technology has been used to solve many different problems including object detection, object tracking, navigation, and meteorological sensing. More recently, radar has also been used to detect objects underground, a technology commonly referred to as GPR. For a comprehensive overview of GPR's history, read [1].

Many types of systems have been constructed to use remote sensing technologies to model volumes underground. There are many potentially useful remote sensing technologies including radar, Electromagnetic Induction (EMI), infrared, and acoustic sensors. The major advantages of radar as compared with other sensing technologies

include the range over which it can operate, the ability to see through multiple obstacles, the ability to detect multiple targets, and the ability to detect many different types of materials. Radar systems often cannot have all of these capabilities at once, they must be optimized for certain properties over others. For a given level of broadcasting power, GPR generally faces one major tradeoff, the scale of object it is capable of detecting--as small as millimeters--or as large as meters--and the depth into the ground it is capable of penetrating. These two properties are a function of the waveform emitted, the power of the radar system, and the dielectric properties of the materials through which the waveform propagates. This means that different GPRs are ideal for different purposes. Utility workers may generally care about very small objects close to the surface, while a geologist may be interested in rock layer densities as deep into the ground as can be sensed.

Regardless of the specifics of the system, all GPR systems work in approximately the same manner. A signal is generated and sent out the transmitting antenna. The signal propagates through the air and into the ground. Every time the signal passes through a change in impedance, a copy of the signal is reflected back. Changes in impedance are experienced by the wave whenever the medium's dielectric constant changes. Returning signals are received by the receiver antennas in the GPR system. Since the signal travels at a known speed, by timing how long it took to receive the signal, the distance to the change in material can be calculated. The magnitude of the received signal is affected by several variables, most notably the radar cross section of the object. Using multiple transmitter and receiver antennas enables more sophisticated techniques. Many GPR systems use a number of antennas to cover a wider area.

Turning the raw data from a radar's receiving antennae into a meaningful understanding of the world is not an easy task. Before this can be accomplished, artifacts from the sensing process must be removed. Spurious signals from components in the radar circuitry, background radiation, and signals originating from known equipment, collectively called noise, must be eliminated by a combination of hardware and software processing techniques. Once preprocessing and noise elimination has been conducted,

the process of separating relevant objects from unimportant objects, called clutter, can begin. Once clutter has been removed, the resulting identified objects are the end product of the GPR system. Noise and clutter processing is still an ongoing subject of research. While aircraft identification has been studied for decades for airborne radar, GPR still employs human signals analysts instead of trusting an entirely automated system.

One approach to creating better object detection systems is to compare models of the same place taken at two different times. This approach is called Change Detection (CD) and it provides several desirable properties. If one can assume that objects present in both scans are uninteresting, then the challenge of finding interesting objects can be simplified.

Using the CD paradigm is not without added drawbacks. Many changes are due to irrelevant processes, such as environmental conditions, wildlife, or simply the passage of time. Increasing the sensitivity of a remote sensing system decreases the chances that small changes will be missed, but it also generates more uninteresting changes. Consider the case of a system that can sense a centimeter of alteration in the height of dirt under it. This capability will raise the chance of detecting a place where someone buried an object by detecting the disturbed earth around the object. This capability will also also create unhelpful false positive signals due to vehicle tracks, puddles, or a dog's paw prints. Lessening the sensitivity of this system would resolve some of these unwanted scenarios, but it would also hide a very carefully buried object. Therefore there is a tradeoff between the desire to make more sensitive systems to increase the chance of detecting useful signals and the desire to make less sensitive systems to keep the output useful.

GPR devices can be built in many formats including handheld, small portable wheeled devices, vehicle-mounted, or attached to aircraft. These systems collect data in different manners depending on the size of the area they can cover and the manner in which they are moved through space to sweep out coverage. Each of these different categories of system present different challenges to provide effective coverage. Systems with less automation controlling their movement, such as handheld devices,

require strict adherence to proper procedure to avoid creating gaps in sensor coverage [2]. More automated systems such as Drive-By-Wire (DBW) vehicles can sweep out more uniform patterns, but suffer from added movement constraints and challenges associated with driving a vehicle. This thesis will focus on the challenges of using a vehicle-mounted ground-penetrating radar systems with drive-by-wire capabilities.

1.2 Gopher

Gopher is an experimental vehicle-mounted GPR system developed in partnership with MIT Lincoln Laboratory. The goal of the Gopher system is to allow an operator to find buried objects. The system can operate in several ways, and the one studied here uses change detection to allow an operator to combine data from multiple passes over the same place. Gopher consists of a stock vehicle, GPR array, computer system, drive-by-wire system, and a special purpose display screen. Gopher operates by using a drive-by-wire system to drive a previously-traversed route. Gopher's computer system constantly compares the data coming from the GPR array to previously recorded GPR data. When a difference above a certain threshold is detected, the drive-by-wire system halts the forward motion of the vehicle and displays the GPR data to the operator for an analysis. Further details of this process are discussed in Chapter 3.

Gopher faces a unique challenge compared with previous GPR systems. Gopher is intended to operate on a wide variety of potential object types, shapes, and sizes. Normally, object detection problems are addressed by building a library of known objects. By predicting how an object might appear on radar, unknown signals can be matched against a library of known objects to determine the unknown object's identity. This process works well in environments in which it is easy to identify and discard uninteresting signals or where it easy to identify all possible interesting objects for a particular application. Prior work in humanitarian demining provides a rich background of resources that addresses these cases where a specific class of targets is known (see Chapter 2). It is substantially harder to accomplish the same task when uninteresting signals and interesting signals look similar. It is even more challenging when there is no precise definition of an interesting object, as is the case for Gopher.

1.3 Motivation

Users of GPR include utility companies, humanitarian demining operations, surveyors, archaeologists, and environmental scientists of many varieties. These users all rely on radar's ability to conveniently and efficiently sense objects remotely, safe from potential harm, and provide increased situational awareness to human operators with mission objectives.

Understanding what is beneath the surface of the ground is useful for countless purposes and applications. The simplest way to obtain this information would be to excavate the area, but this is frequently not a practical or desirable option. Sometimes it is physically impossible to excavate an area without causing damage to nearby structures or because doing so would disrupt normal use of the space. In other situations it is simply too dangerous to dig because of hazardous materials or environmental conditions.

Failure to properly use remote sensing technologies are often grave. Failing to detect objects can have serious monetary and loss of life consequences in many different domains. Utility companies have faced the dangers of digging into city streets ever since pipes, gas, and communications infrastructure began to be buried. The United States is facing a large and costly problem as it confronts its aging buried infrastructure [3]. Damaging a natural gas line can result in large evacuations or explosions [4]. Severing cables accidentally is a common construction mishap, and it can have serious consequences and require very expensive repairs [5].

The consequences of failing to detect explosives can lead to terrible injuries and death. As of 2009, an estimated 45-50 million mines have been left in wartorn areas of the world according to the U.S. State Department [6]. Even assuming that no new mines are put in the ground, at the current rate of removal it is estimated that humanity will still be removing mines for the next 450-500 years [6]. Efforts aimed at reducing casualties from Improvised Explosive Devices (IEDs) in campaigns in Iraq and Afghanistan have shed new light on the growing prevalence of IEDs as a worldwide threat. According to Army Lt. Gen. Michael D. Barbero, director of the

U.S. Joint IED Defeat Organization, between January, 2011 and September, 2012, there were more than 10,000 global IED events in 112 countries [7].

In addition to disaster avoidance, there are also highly beneficial uses of remote sensing technologies to study the environment. Natural resources companies have employed many different technologies to help find oil, natural gas, and even mineral deposits [8, 9]. Scientists frequently rely on remote sensing technologies for large geologic surveys, helping to map out the structure of the subterranean world [10, 11]. Recent oil spills have drawn attention to the use of GPR to map contaminants [12, 13].

1.4 Human Factors

GPR systems used can be used for many purposes, and the efficacy of every use depends upon the ability of the system to differentiate relevant signals from noise and clutter. As a formal evaluation of the Gopher system will conclude in Chapter 3, the central challenge of this research effort is to assess and improve the ability of the system to provide object discernment capabilities. This process can be accomplished in many ways, and it is the central challenge of this research effort.

One approach to improving the capabilities of a detection system is to make it more sensitive. More sensitive radar detection capabilities can provide more complete and more accurate models of the world for decision making. However, this is only true if the automation and human interfaces included in radar systems are appropriately designed to serve their operators' needs. This thesis explores the layers of technology between GPR sensing hardware and human operators with the goal of improving operator situational awareness and decision-making capabilities.

While many systems for traditional object detection have been studied in the past, far less attention has been paid to the human factors of GPR change detection systems. Despite extensive research into automatically mitigating noise and rejecting clutter, human operators ultimately must deal with artifacts that appears in operator displays. This thesis will aim to both document the ability of subjects to differentiate objects from false alarms on a traditional GPR display, and assess how operator performance differs on a novel GPR display.

1.5 Problem Statement

Despite advances in radar technology, signal processing, and artificial intelligence, human operators are employed to process GPR data in many applications. When searching for a wide variety of possibly relevant signals, the detection of objects with automated technologies has not advanced to the point where it is sufficiently flexible and reliable to be preferable over human judgement. Therefore investigating the ability of human operators to distinguish objects from noise is important to the advancement of GPR-based object detection systems. In addition, modeling human behavior in this domain provides insights into how automation could assist a human. Understanding the currently achievable levels of performance and the relative merits of both human and automation elements enables engineers to construct the most functional systems possible.

GPR detection systems do not operate in isolation; they relate to an ecosystem of other systems which collectively provide various capabilities. Understanding how all of these systems are used and how the operator interacts with each of them is key to designing each component, especially the human user interfaces for each component. The primary area of focus for evaluating the system's ability to provide signal detection capabilities is its ability to support the operator in differentiating signals from noise. The best possible GPR detection system would enable operators to efficiently, correctly, and reliably find objects of interest using a minimum of the operator's attention. This research aims to design and test a new interface's ability to meet these goals.

1.6 Research Objectives

The scientific method suggests that to find solutions to the problem statement, the abstract goals of this research effort should be to:

1. Understand and model the phenomena involved.

2. Construct a hypothesis that addresses the problem.

3. Evaluate the hypothesis experimentally.

4. Document the results and suggest further steps to address the problem.

These goals were translated into specific research objectives with concrete outcomes:

- *Objective 1*: Determine the functional and information requirements for a vehicle-based change detection GPR signal analysis tool.

- *Objective 2*: Design an improved interface which supports human analysis of GPR signals.

- *Objective 3*: Evaluate the effectiveness of the interface design through experimentation and the use of credible metrics.

1.7 Thesis Organization

These objectives above were accomplished and documented in this thesis. The full details of this effort are presented in six chapters.

- *Chapter 1: Introduction* presents the motivation for this research and a summary of the goals this thesis hopes to address.

- *Chapter 2: Background* summarizes topics relevant to this research from various disciplines and related fields.

- *Chapter 3: Interface Design* explains the human factors analysis process used in this research and the resulting interface design.

- *Chapter 4: Usability Analysis Experiment* describes an experiment design using human subjects evaluating the interface design from Chapter 3.

- *Chapter 5: Results and Discussion* examines the data and statistics calculated from the usability evaluation Experiment.

- *Chapter 6: Conclusions* discusses the implications of the Results chapter and suggests topics for further research.

Chapter 2

Background

2.1 Demining

2.1.1 Relevance

While many applications of GPR and similar sensing technologies have been studied, demining is a particularly useful example. Demining has benefited from decades of study by military organizations, and therefore there is a rich body of related research quantifying system performance and suggesting further research topics. Some of the similar challenges between demining and generic object detection which make demining research relevant include:

- Sensor physics and hardware design for many sensor types
- Localization and registration of signals
- Preprocessing of sensor data into a model of the environment
- Noise rejection
- Clutter rejection
- Visualization of data for human operators
- Machine learning and AI development for automated detection
- Training procedure development
- Operation procedure development for human decision making

2.1.2 History

The trends in both sensor and interface design through the history of demining technologies illustrate the leaps that have improved the capabilities provided by subterranean detection systems. The evolution of remote sensing of mines starts with single-operator handheld detector devices. Demining devices are usually optimized for finding metallic buried objects up to a few centimeters into the ground in areas that contain vegetation and clutter. Mine detection devices have employed many different sensing technologies, one of the most common being EMI, the technology behind metal detectors. Starting in 1944, some manufacturers started to replace metal components with nonmetallic materials, which made EMI sensors less effective. Within the last several decades, a new approach to detecting low-metal mines has been developed by employing hybrid sensor systems that add GPR to existing EMI systems. These hybrids add sensitivity and improve clutter rejection, particularly in low-metal mine situations [14].

Ignoring physical risk, the biggest problem with handheld demining devices is that they must be used properly by operators or they are useless. Proper operating technique has many aspects including keeping the device level and moving it very slowly forward. One of the hardest problems is sweeping the device over the lane correctly to avoid missing any part of the area. Handheld devices typically provide sensing capability straight downwards in a narrow (high-gain) directional pattern. Therefore it is essential that operators attempting to clear a large area move the device in a systematic fashion that guarantees complete coverage. Patterns that are used to completely cover a large area require a several centimeter grid of sweeping motions. Achieving this level of fine coverage is slow and difficult. A common topic of research in demining programs attempts to address the issue of gaps that can be created by operators improperly sweeping an area [15, 16].

Many human factors engineering solutions have tried to find better alternatives to trusting operators to sweep correctly. One area of research attempts to address gaps and coverage problems by automating the task of moving the same sensors used

in handheld devices back and forth. Several experimental robotic systems have been invented to improve upon human reliability [17, 18]. These systems have the benefits of automated consistency and the ability to repeat a boring task with little deviation without tiring a human operator. These systems also take a step towards removing the human operator from a dangerous environment.

Other advances in detection technologies have provided additional area coverage with one device. Vehicle mounted sensing systems, for instance, can provide a line neighboring of sensors that can be moved perpendicularly to their long axis to sweep out a volume of coverage. This type of platform can address many of the problems found when covering a large area of terrain by avoiding the creation of small gaps. This technique assumes that the area is smaller than the width of the array. If multiple passes are needed to cover the area, then the gap problem still exists, except on the scale of meters instead of centimeters. A further extension of this trend are aerial detection systems, which can cover even greater areas [19].

2.1.3 Mine Detection Techniques

Regardless of the technology employed, all sensing systems must either convey information to a human operator or trust an entirely automated system to detect mines. Many handheld systems use acoustic representations of the sensor data to give a human operator an interactive way to explore underground. Developing better acoustic representations is an area of active human factors research [20, 21, 22, 23, 24]. Newer demining systems often feature multiple sensor types and combine sensor data into a visual representation of potential targets. More recently, these visualizations have started to include algorithmic assistance to detect mines, though these systems are still experimental [17].

If signal profiles of known targets such as mines are available for comparison, a library of relevant signals can be compared against a received signal to classify the signal as a mine or noise. This is the key point which differentiates demining technologies from generic object detection systems. Mines are highly regular objects with a known geometry and material composition. Attempting to look for highly

specific object signatures is a tremendous simplification of the generic object detection problem. In the general detection problem, systems must be designed to detect any type of buried object, and therefore they cannot rely on matching known signals.

One aspect of demining which receives special attention in research efforts is the process of training new operators. There is clear evidence that experienced handheld operators perform better than novice operators [25, 26]. As with any research which involves learning, this complicates experimental design for studying human performance because subjects must have similar levels of experience for the results of the experiment to be meaningful. Prior literature on training is most relevant to this research effort because developing better interfaces for training can lead to better interfaces in standard equipment.

2.2 Decision Theory & Classification

A simple model of a classifier can be thought of as an algorithm applied to a stimulus as illustrated in Figure 2-1. From this diagram we can see that there are three sources of complexity in the classifier process: the nature of the phenomenon, the properties of the sensor, and the classifier algorithm's behavior.

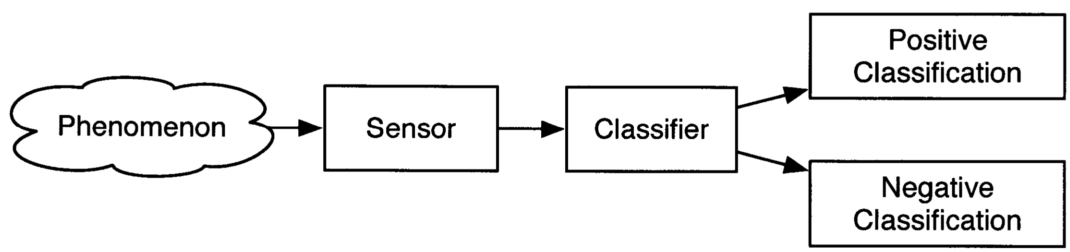


Figure 2-1: Binary Decision Model with a Sensor

The simplest decision algorithm is a threshold, and many other algorithms are derived from a threshold applied to a complex metric. If the stimulus is stronger than the threshold, then one result is returned; if the stimulus is weaker than the threshold, then the other result is returned. If the range of stimulus values from a true signal is entirely distinct from the range of stimulus values from a false signal, then a threshold

can achieve perfect performance as illustrated in Figure 2-2.

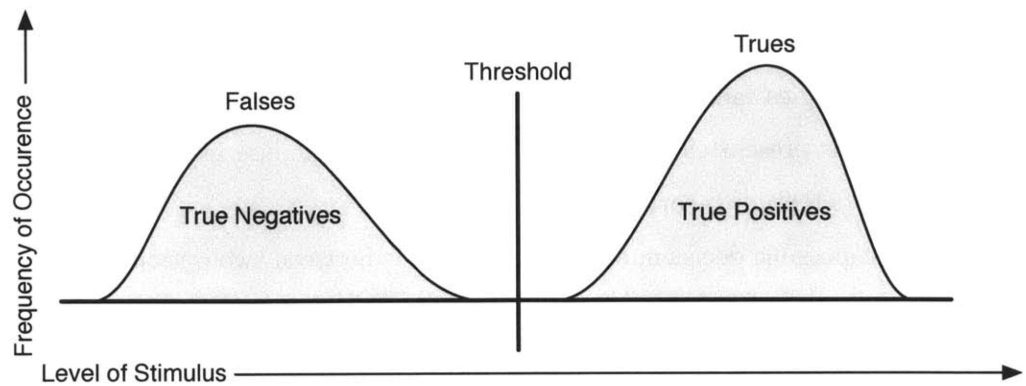


Figure 2-2: Binary Decision Threshold on Non-Overlapping Stimuli

In nearly all applications this ideal scenario does not reflect reality due to sensor noise or the underlying phenomenon's properties. By choosing the stimulus level for the threshold, a trade-off can be made between the risk of a false positive and a false negative as seen in Figure 2-3. As the threshold is changed, a different percentage of each case will be misclassified.

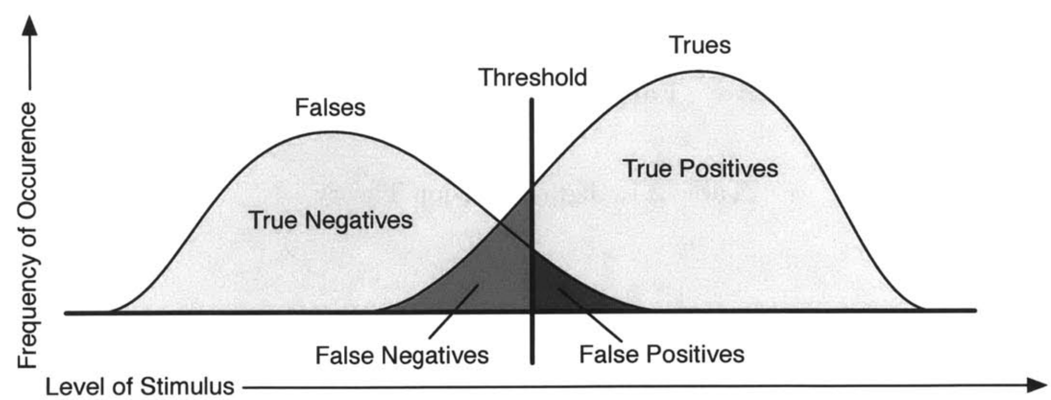


Figure 2-3: Binary Decision Threshold on Overlapping Stimuli

The performance of a classifier is therefore dependent not only on the decision process, but also on the nature of the stimulus being classified. For some collections of

signals, many classifiers may achieve identical performance, while the same classifiers may have differing performance when given other stimuli.

The identification of useful signals from noise is not a challenge unique to GPR. All sensor technologies face a similar hurdle. The exercise of deciding whether an interesting signal is present can be studied as an abstract exercise in decision making known as signal detection theory.

The simplest possible decision, a binary decision, is between two options. In signal detection theory, the two choices are "signal present" (also known as positive) and "signal absent" (negative). Signal detection can also be thought of as a classification problem because many decisions can be thought of as attempts to classify an object into one of two categories.

Every binary decision can be thought of in terms of two dimensions: the oracle 'truth' of the situation (either true or false), and the answer provided by the classifier (either true or false). For any instance of a binary classification decision there are four possible results depending on the combination of these variables summarized in Table 2.1.

		Oracle Truth	
		True	False
Decision	Positive	**True Positive** "Hit"	**False Positive** "False Alarm"
	Negative	**False Negative** "Miss"	**True Negative** "Correct Rejection"

Table 2.1: Signal Decision Theory

2.2.1 Classifier Performance Metrics

The performance metrics of a classifier can be mathematically calculated by examining the frequency with which the four possible outcomes from Table 2.1 occur as a fraction of the total number of scenarios. For sensor research, these numbers are sometimes calculated per unit of area covered instead of the total number of scenarios. An ideal classifier would only yield true positives and true negatives, collectively called

'correct' answers. The worst possible classifier would only return false positive and false negatives, 'incorrect' answers. Most classifiers are somewhere between these two extremes, returning some number of correct and incorrect decisions.

While correctness is perhaps the most relevant performance characteristic of a classifier, there are other dimensions of classifier performance. In most applications, a false positive and a false negative are not equally preferable. Several metrics can be used to analyze the performance of a classifier with respect to the false positive versus false negative trade-off. The two most common metrics are the True Positive Rate (TPR) and False Positive Rate (FPR) calculated as indicated in Table 2.2. A detector with a high TPR will rarely miss a positive signal. The FPR represents the probability that a signal will be classified as positive given that the stimulus is *not* present. A classifier with a low FPR will only detect real objects and will not waste the operator's time with false positives.

TPR is sometimes known as the Probability of Detection (PD) because it represents the probability that a signal will be classified as positive given that the stimulus is present. This usage can lead to linguistic confusion. PD is frequently used by demining literature to refer to the probability of detecting a mine with a given system. This is not the same thing as TPR. The two quantities are related by the target density, the number of targets found in units of area or linear distance. This research analyzed the behavior of the system with an artificial target density and thus does not represent any actual application, environmental condition, or target variety. Similarly, for the FPR, there is no particular area or distance this system covered to generate false detections. (The true-to-false ratio in this research was defined to be 50%, so it is not possible to assess this system in units such as PD% or False Alarm (FA) per kilometer from the data presented.)

2.2.2 ROC Curves

The optimization of the threshold for a classifier begins with a policy decision. This policy decision must make a trade-off between the risks of a false positive and a false negative. For a given set of stimulus inputs and a classification algorithm, the

trade-off between false positive and false negative risk can be visualized with a graphic known as a Receiver Operating Characteristic (ROC) curve. A set of example ROC curves can be found in Figure 2-4. Such a figure is generated by choosing a range of threshold values and plotting the TPR and FPR values for each possible threshold. ROC curves allow policy-makers to study the trade-off between false positives and false negatives for a particular sensor. The relationship between the ROC curve and the sensor model from Figure 2-3 is useful to examine. If the risks of false positives and false negatives are considered equal, then the logical choice of decision threshold is the one that minimizes the total number of incorrect answers. Such a decision threshold would be found by minimizing the sum of the areas in the false negative and false positive regions of Figure 2-3 and correspond to the point on the ROC curve that is furthest to the top left.

In addition to visualizing the tradeoffs for a single sensor, ROC curves can also be used to compare different sensors. A sensor with an ROC curve further towards the top left corner of the plot is strictly superior in performance to another sensor with an ROC curve further towards the bottom right. In most cases such comparisons are possible, though different sensors may have crossing ROC curves or curves which asymptotically approach the same limit in certain regions of behavior.

It is important to remember that ROC curves can be deceptive if plotted with a line rather than a series of points. Only some values on an ROC curve can actually be obtained simply by the correct choice of a threshold value. However, an advanced

Abbreviation	Metric Description	Definition
height TP	Total true positive count	Observed from data
FP	Total false positive count	Observed from data
TN	Total true negative count	Observed from data
FN	Total false negative count	Observed from data
P	Total positive	TP + FN
N	Total negative	TN + FP
TPR	True positive rate	TP/P
FPR	False positive rate	FP/N

Table 2.2: Binary Decision Metrics and Formulas

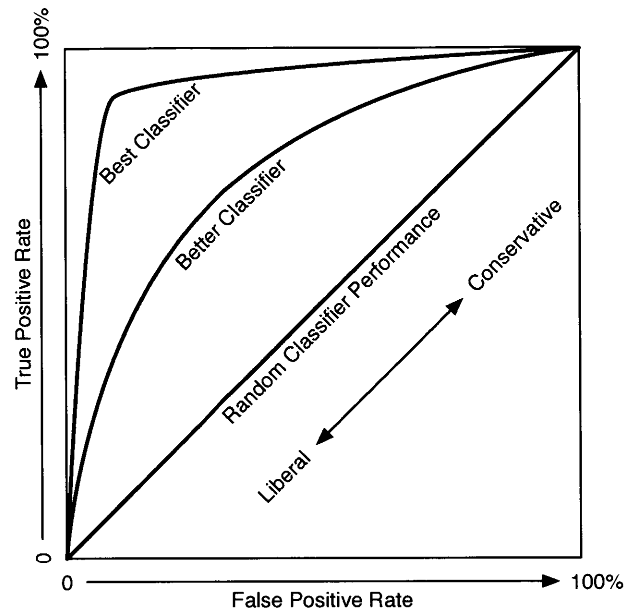


Figure 2-4: Example Receiver Operating Characteristic Curves

technique can be used to obtain any point in the convex hull of an ROC curve. By choosing any two points in the convex hull of the curve and randomly choosing between them with proportion p, the points along the line between them can be obtained depending on the value of p.

2.2.3 Timing

In addition to the quality of the classifier's output, many applications demand time constraints on their classification algorithms. Timing for a classifier is often measured by two metrics: the time it takes to train or preprocess the library of training scenarios, and the time it takes to classify a new scenario. For most real applications of artificial intelligence, we can assume that an arbitrarily large amount of time could be used to train a classifier, since the time is spent only once. Therefore only the time it takes to classify a new scenario matters for artificial intelligence classifiers classifiers.

2.3 Interfaces for GPR Signal Detection

Many early systems using novel sensor types employ a human operator to accomplish the signal detection task. Early detection of submarines in the North Atlantic Ocean by the SOSUS system, for instance, was accomplished by human analysis of acoustic spectrum plots [27]. As mentioned earlier in the discussion of demining interfaces, many EMI and EMI/GPR hybrid devices use an acoustic interface where the properties of objects are relayed to an operator using volume and pitch. Several different types of visual displays have been invented to display radar data to users for tracking aircraft and naval vessels [28].

GPR analysis for survey purposes is still largely conducted by humans. Early systems plotted signal strength against time in a graphical readout of the sensor values on waterfall plots as shown in Figure 2-5. These individual waterfalls visualize an "A-scan" by showing a signal deviating from vertical. When multipe A-scans are presented next to each other, these plots are called "B-scans" and usually represent a history of where the sensor has travelled over. If successive measurements were taken with even temporal spacing, or the measurements were taken at a precise distance from the previous measurement, one can adapt this data into a spatially organized B-scan. It is common to replace the horizontal deviation from a waterfall display with a gradient. See Figure 2-6 for an example B-scan rendered using a gradient with suspected objects marked. Since one can convert between distance and time using the speed of the wave propagation, these scans may use distance or time for the vertical axis. "C-scan" data is accumulated by assembling multiple B-scans. Typically C-scans are visualized by taking a slice of the data perpendicular to the motion of travel of a multi-element radar array (perpendicular to a B-scan). However, B-scans and C-scans present slices of the underground, and therefore present a limited view into a volume.

Regardless of the specific display used, current GPR systems show the GPR output to a human, and the human is responsible for analyzing the imagery and generating a mental model required to formulate a plan of action. At least one experimental GPR

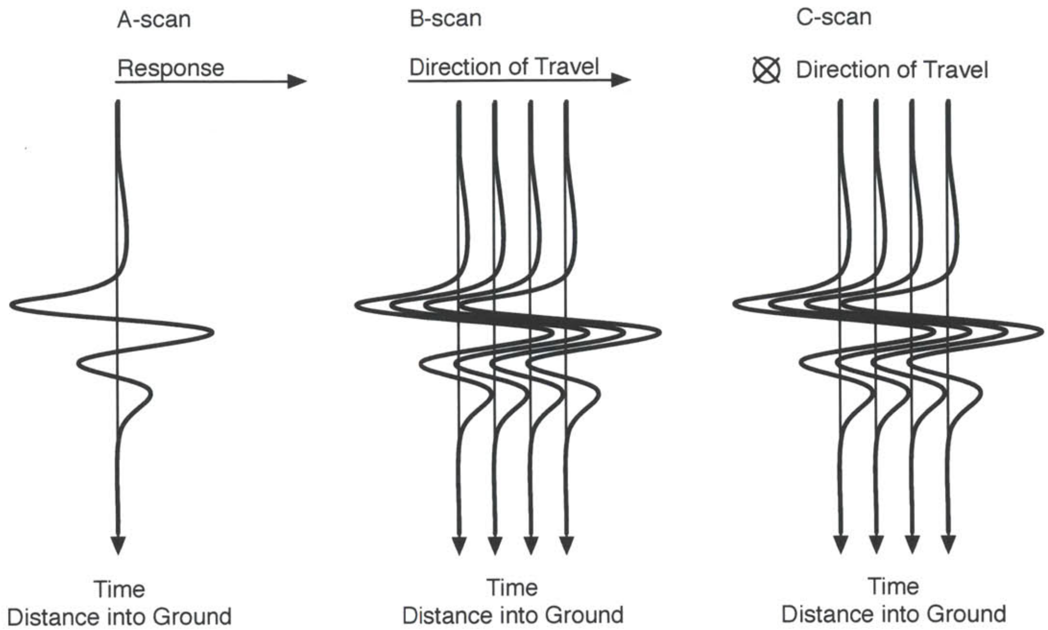


Figure 2-5: GPR A-scan, B-scan, and C-scan Concepts

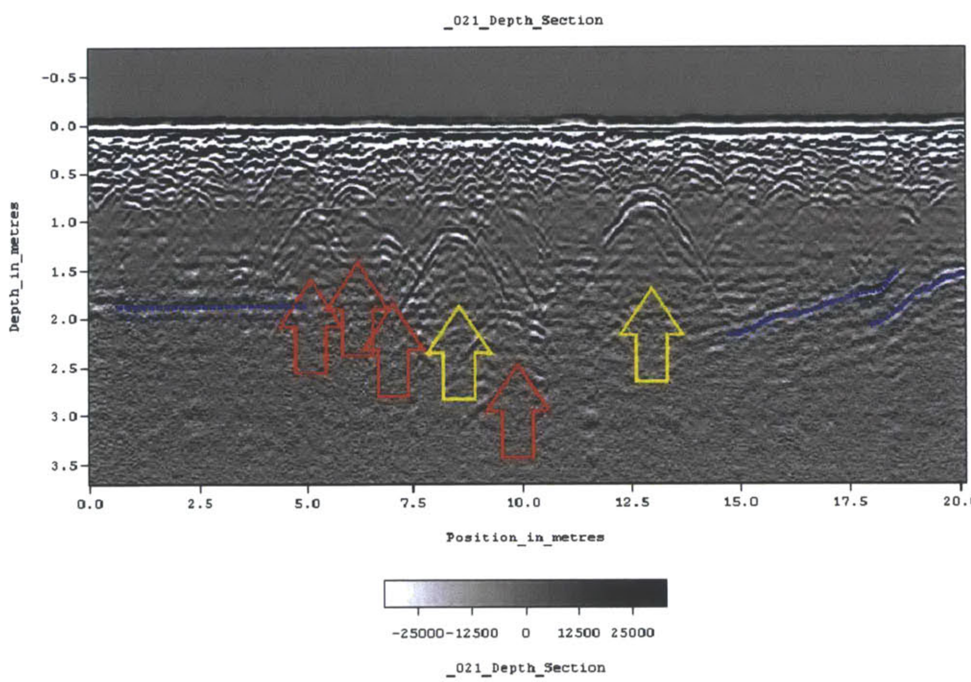


Figure 2-6: Example GPR B-scan Rendering of a Cemetery

Accessed from http://upload.wikimedia.org/wikipedia/commons/9/9c/LINE21.jpg
Image released into the public domain.

system, RMS, has presented images to operators in a manner that shows C-scan data from the top [17].

2.4 Summary

This chapter builds on Chapter 1, presenting information on an area of research closely related to Gopher, demining. In addition, background information on signal detection theory was discussed in preparation for the analysis of the experiment found in Chapter 4 and Chapter 5. Finally, a summary of various displays for GPR was presented to cover the basics of how GPR data has been traditionally displayed. The next chapter begins the human factors evaluation of the Gopher system and suggest a new interface design based on all of of the background information presented thus far.

Chapter 3

Interface Design

3.1 Overview

This chapter presents a human factors analysis of the Gopher system and the design rationale for a new interface. Since Gopher was not a production system at the time of this research, the analysis here is based on the state of the system design late in the implementation process. The pool of available subject matter experts for this system included the project leader, the operators who used the system during testing, and the engineers implementing the system, rather than a traditional end user population.

The stages of the human factors analysis and design process used in this work were as follows:

1. Develop scenario task overview of the system's operation

2. Document a processes model for the system.

3. Identify the functional and information requirements of the system.

4. Develop an improved design for the future system.

5. Implement the new design.

6. Test the old and new designs with human subjects.

7. Analyze the differences between the designs.

Steps 1--3 were conducted using the hybrid Cognitive Task Analysis (hCTA) process elaborated in this chapter. An improved design and a supporting rationale is presented at the conclusion of this chapter. The experimental process for analyzing the new interface is presented in Chapter 4 and the results are discussed in Chapter 5.

3.2 hCTA

An hCTA [29] was performed to observe and document the Gopher system. A series of interviews was conducted with the project leader and two operators who operated the system during its testing. In addition, direct observations of the system in use during testing were made. From these observations, a scenario task overview and event flow model were constructed.

The scenario task overview found three phases of operation:

1. Route Selection: The operator selects a route to follow and activates the automation.

2. Primary Operation: The operator allows the automation to drive the vehicle while he or she monitors the environment.

3. Signal Analysis: The operator responds to potential objects found underground by analyzing the GPR data and engaging external actors if neccesary.

A typical period of operation involves a route selection phase, followed by any number of primary operation and signal analysis phases.

An event flow model of this scenario was constructed to understand the system in typical operation. A graphical representation of the event flow model is shown in Figure 3-1.

Table 3.1 presents the functional and information requirements for the Gopher GPR system. These requirements answer the fundamental questions:

1. What actions must the operator be able to perform with the Gopher system?

2. What information must the operator be shown by Gopher?

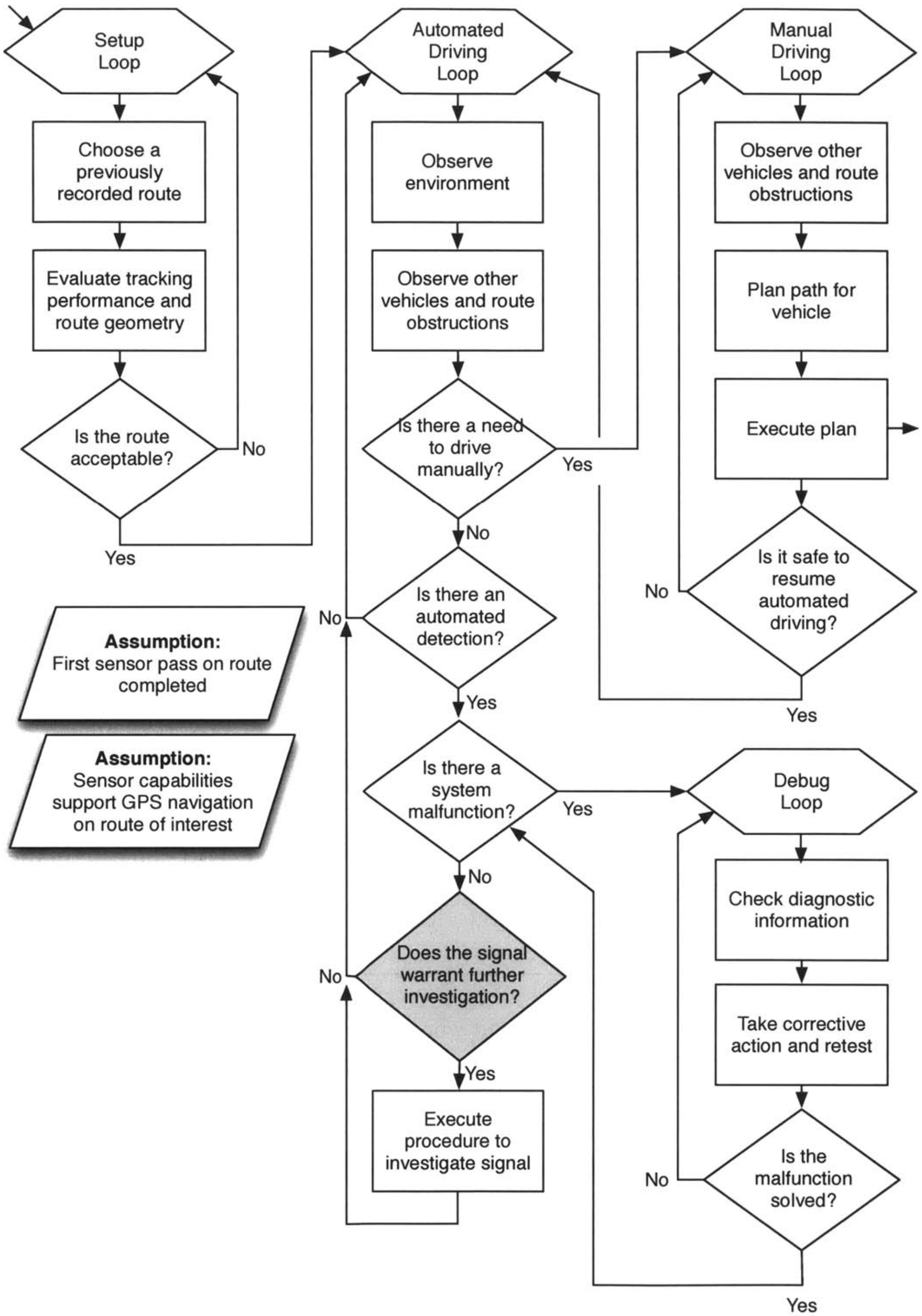


Figure 3-1: Gopher Event Flow Model

Table 3.1: Gopher Functional and Information Requirements

Functional Requirements
Activate the auto-brake The operator must be able to activate the brake quickly before driving over an object. This is manually possible using the DBW system or physically stepping on the brake. There is also a software capability which would allow automation to trigger a braking action. In normal operation, the operator leaves the job of braking to an automated subsystem.
Deactivate the auto-brake The operator must be able to override any automation controlling braking and operate the vehicle in the event that he or she disagrees with the system's automation. This process is handled using an override button external to the Gopher interface hardware. The automation could render the vehicle unable to move if no override capability existed.
Report the location of a buried object The operator must be able to inform an external person that he or she has found something that warrants further action. This is accomplished outside the Gopher hardware with radio communication.

Information Requirements
Vehicle path The operator must be able to understand the current position of the vehicle relative to the previous sensor pass. The system only provides effective detection capabilities in the areas with overlap between the two passes. The DBW navigation does not perfectly match the previous traversal at all times, and the operator must be able to detect a problem if it begins to occur.
Location confidence The operator must be able to understand when the vehicle fails to know its position accurately. Failing to properly register the vehicle's position will results in excessive signal generation.
System status The operator must be able to understand when system components fail or are turned off. This information is available in the form of binary or trinary status indicators from various components that self-analyze their own status.
Object detection The operator must understand when an object is under the sensor. In order to do this, the operator must be able to understand various properties of the signals including the size, location, shape, and edge gradients.

3.3 Research Focus

This research effort aimed to bring about the largest impact possible with limited resources. Due to its central role in the event flow model, the signal analysis process (see the gray diamond in Figure 3-1) and GPR display were determined to be the most crucial process and corresponding interface in the Gopher system. In addition, some components of the Gopher system, such as the stock vehicle, are substantially harder to modify than the new hardware introduced for Gopher's GPR processing. Therefore, while there are many other aspects of the Gopher which would benefit tremendously from a human factors improvement process, the GPR display was chosen as the target of attention and detailed analysis.

In addition, it is useful to remember that the automation is performing a level of decision making before the human is presented with a decision. As a consequence, there are two varieties of potential errors, those made by the automation, and those made by the human. This thesis does not address the category of errors made by the automation, which would need to be addressed through modifications to the radar itself.

3.4 Old Interface

Although the Gopher system was not in production use at the time of this research, an existing interface concept had been developed. The old interface displays GPR C-scan data to the operator using a heat map by summing in the z-dimension and organizing the data by radar element and time. Appendix A contains the slides used to train experiment subjects, and the interface is depicted in Figures 3-3, 3-2, 3-4, and 3-5a.

The interface concept at the time of this research used a full rainbow color scheme to represent radar data. For experimental consistency, the reference implementation of the old display (as seen in Figure 3-5a) uses the color scheme of the new interface. See Figure 3-7 for a comparison of the colormaps.

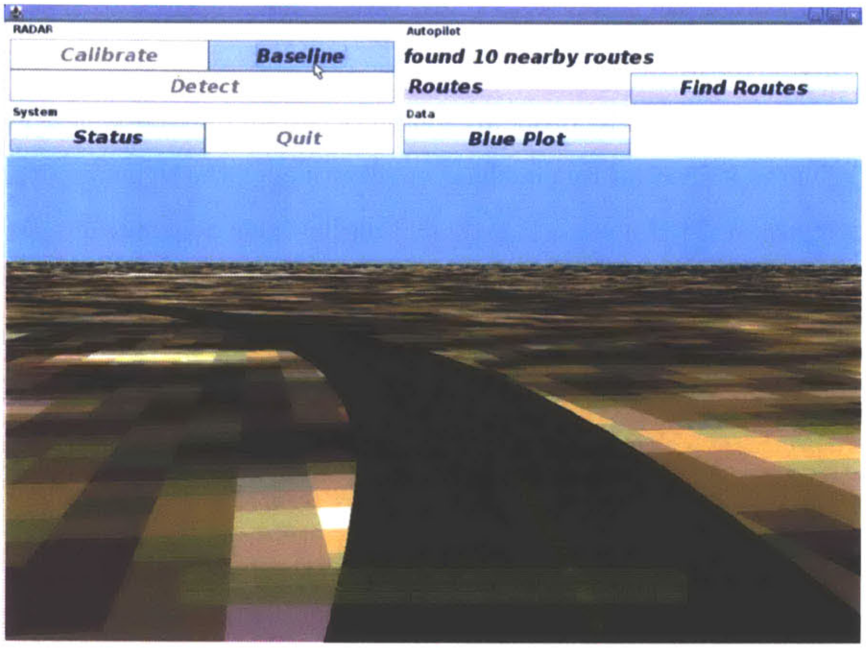


Figure 3-2: Old Gopher Interface: Normal Operation

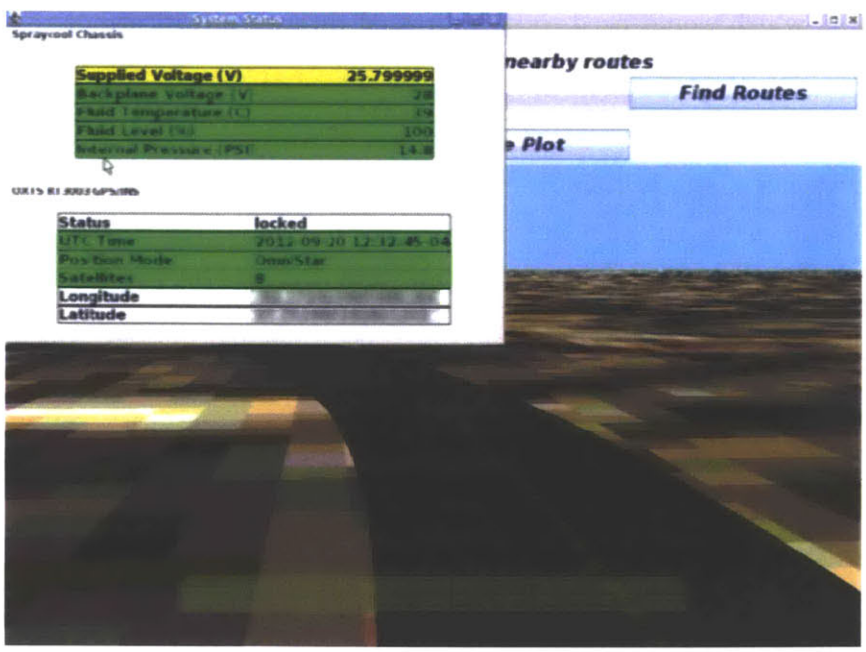


Figure 3-3: Old Gopher Interface: Viewing Status Information

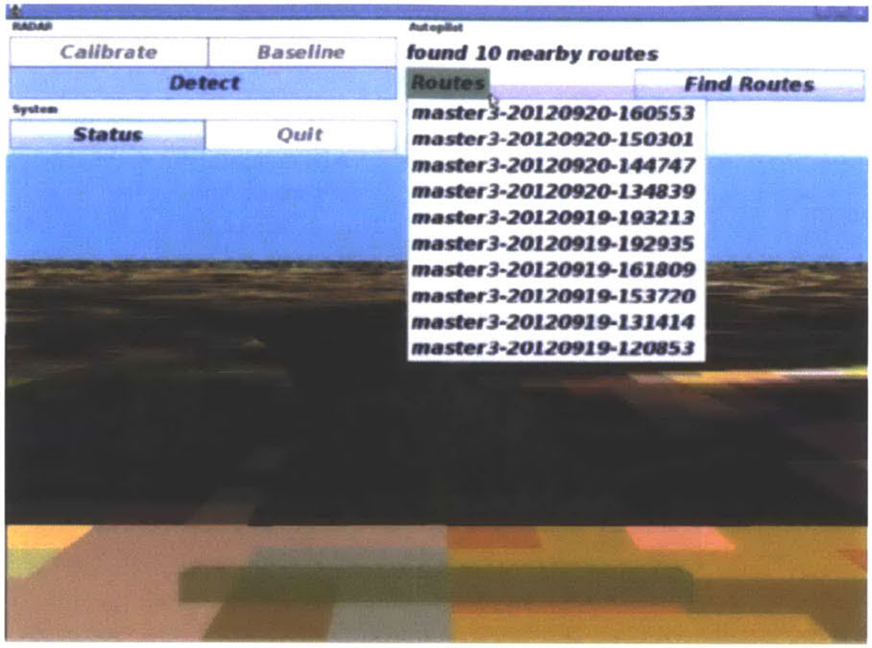


Figure 3-4: Old Gopher Interface: Selecting a Route

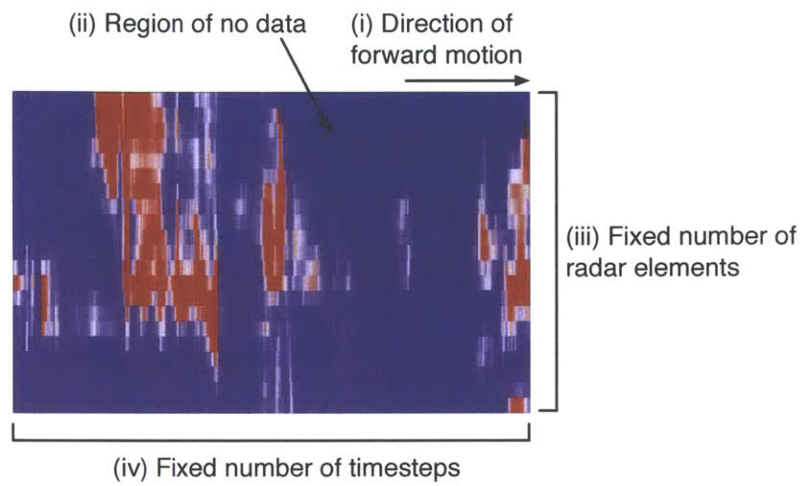
3.5 New Display Design

A new GPR display was designed to meet these requirements. Figure 3-5b shows an example of the old and new displays with annotated features. The individual features are discussed in the following subsections.

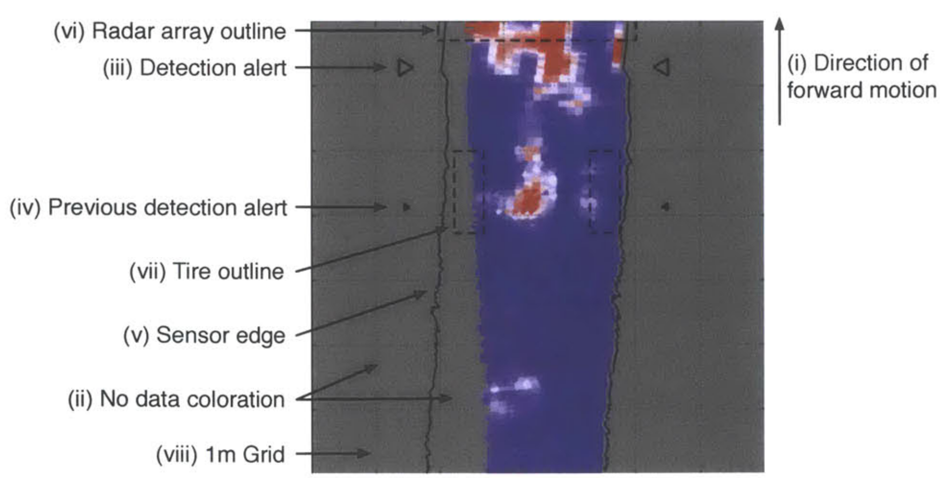
3.5.1 Spatial Display of Data

The Gopher radar data is captured and analyzed by various algorithms using the location of each sensor at the time the data is recorded. In the old interface, (see Figure 3-5a), it is displayed to the user with a colored display of rectangles, organized by time and sensor number (Figure 3-5a (iii)). A fixed number of sensor readings is displayed at all times, with older values being removed from the left edge to make room for new readings added to the right edge (Figure 3-5a (iv)).

Since the data is captured on a polling timer, this means that the graphical representations of objects are distorted on the display depending on how quickly the sensor is traversing the ground. Objects are also distorted by the turning of the



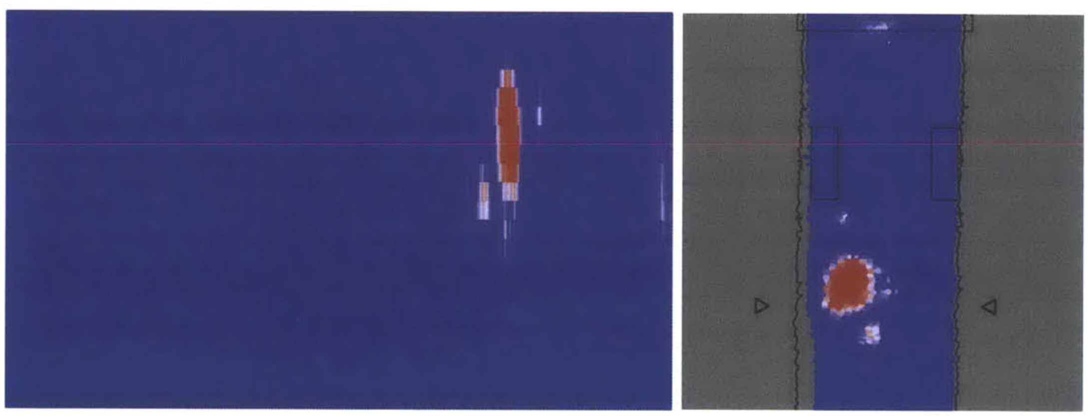
(a) Old Gopher GPR Display: Viewing a Detection



(b) New Gopher GPR Display: Viewing a Detection

vehicle, which decreases the temporal spacing between sensor readings on the inside of the turn compared to the outside of the turn. Measurements will not be taken if the vehicle has moved less than a specified small distance since the last measurement set. This helps maintain some level of spatial regularity in measurements by ignoring new data when the vehicle is stationary or backing up, though it introduces complex behavior in the location of the individual measurements.

In addition to distortion caused by the location of the measurements, there is also distortion because the aspect ratio of the rectangular blocks of this display do not render objects with the same aspect ratio as they have in reality. A circular object, for instance, would be highly compressed in the horizontal direction and appear as an ellipse with the long axis running vertically. See Figure 3-6 for a comparison of a circular signal on both displays.



(a) Old Interface, Object

(b) New Interface, Object

Figure 3-6: Comparison a Circular Signal on Both Interfaces

In order to improve the ability of operators to understand where a signal is located geographically, the new display renders the data as points on a map. This is accomplished by using a nearest-neighbor interpolation of the radar data which has been geolocated and rotated on a map. A nearest-neighbor interpolation was chosen over other continuous interpolation techniques to avoid removing key information about the gradient of the data.

3.5.2 Track-Up Display

The old interface displays the newest data on the right edge of the display (Figure 3-5a (i)). This is called a "track-right" display and presents the viewer with the experience of traveling to the right. This presentation of the data is at odds with the fact that the vehicle is moving forward, corresponding to the top edge of the display.

To present the data in a more natural fashion, the new interface uses a "track-up" display with the newest data at the top (Figure 3-5b, (i)). This presents the viewer with the experience of driving upwards on the display, similar to how the view out the windshield moves forward.

3.5.3 Data Colormap

The original colormap for Gopher was a full rainbow. Matlab® users would identify it as the default "jet" colormap. This set of colors is unfriendly to color blind individuals. See Figure 3-7 for an example of this colormap. This colormap was chosen because it had superior contrast to standard black and white colormaps. Previous research has favored this colormap over other alternatives reasonably standard in GPR systems [30].

The new interface uses a simpler colormap, based on three colors, a blue, a light gray, and a red. A dark gray is used to signify a lack of data (Figure 3-5b, (ii)). This colormap was chosen for several reasons. It fulfilled the requirement of avoiding combinations of colors confused by individuals with deuteranopia and tritanopia color blindness. In addition this colormap preserves the association of hot colors to stimulus and cold colors to normality. The final choice was inspired by the color theory and cartographic work of Mark Harrower and Cynthia Brewer [31]. The Gopher colormap is shown in Figure 3-7. In this figure, the blue is the lowest radar return value shown and the red is the highest radar return value shown; all values more extreme than these are rounded to the extremes of the colormap.

3.5.4 Signal Identifiers

The process of classifying a detection event occurs in several steps. The first step involves looking at a display and figuring out what part of it to classify. This step is

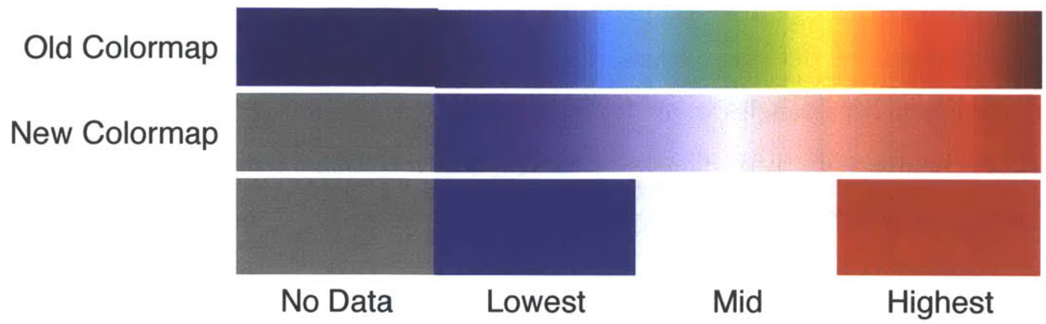


Figure 3-7: Old and New Gopher Colormaps

nontrivial because only a minority of signals are a single isolated peak with no other nonzero values on the display. In the case of a single peak, it is usually obvious what to classify. When multiple signals of various strengths are present in the display then a user must first figure out which one to pay attention to. For instance, signals from earlier detection events are potentially irrelevant, or they may provide examples of false alarms.

The old interface did not assist the operator in this process. Training users involved giving them a general idea what part of the display to look at, without any assistance in a specific case. This is a weakness in the old design because the speed of the vehicle at the moment it begins to brake changes the location of the signal on the display when it comes to rest. If the vehicle moves very slowly, it will brake very quickly, and thus the signal that caused the detection will be very close to the right edge of the display. If the vehicle moves more rapidly, then the vehicle will take a longer time to brake, and thus the signal that caused the detection might be nearly halfway across the display. Furthermore, since the automated system will not trigger another detection until a certain distance has been covered from the last detection, it may not be possible to identify the area on the display that led to the detection event purely by visual inspection.

To help support the operator in identifying the appropriate portion of the data to examine, new indicators were added (Figure 3-5b, (iii)). The indicators consisted

of triangles positioned outside the data region pointing to the potion of the display that triggered the detection event. In addition to triangles pointing to the current detection, smaller triangles were also added to all prior detection events visible on the display (Figure 3-5b, (iv)). This helps operators to maintain a temporal understanding of what has previously been analyzed and what is new.

3.5.5 Lack of Data Indicator

The Gopher system uses multiple passes over the ground to provide data. This means that any location can have zero, one, or two stored passes of information. On the old display, there was no distinction between these three conditions; a zero value was used in places where there was incomplete data. This is highly deceptive because a zero value normally means that the area in question is completely free from objects. If there is incomplete data in a location then the old display gives the operator a false sense of security by portraying that location as free from any potential objects.

In order to help support the operator understand the correct level of coverage, an alternative color, a gray not found in the data color map, was used to indicate a lack of data. In addition, a black line on each side of the path of the vehicle's sensors was added to facilitate understanding the path of the current traversal compared to the reference traversal (Figure 3-5b, (v)).

3.5.6 Vehicle Overlay

When the Gopher system is used to find objects, a secondary process must be used to investigate any objects found. This secondary process might be an additional human with a shovel, or another piece of machinery. In any case, the Gopher system must be moved out of the way in order to allow this secondary process to occur. This means that the operator of the Gopher system must note the the suspected location of the object and convey that information to the secondary process.

The old interface provides no sense of scale or location in the direction of motion. Since the speed of the vehicle at the time of the detection changes the braking distance, the location of a suspected object will be different for different for every detection.

This means that operators must guess the location of the detected object and describe it to another person. Gopher uses a paint spraying system that marks the ground under the corners of the radar array for later reference. However, this does not help the operator figure out how a detection maps to the real world and how to communicate that information to someone else.

To facilitate the operator describing the location of the signal to another person, outlines of some of the vehicle's components were overlaid on the radar display. This provides several landmarks to assist operators. Given the scale of the display, the radar array (Figure 3-5b, (vi)) and the two front tires (Figure 3-5b, (vii)) were visible. In addition to the vehicle components, a faint grid overlay was added to provide a scale (Figure 3-5b, (viii)).

3.5.7 Screen Real Estate Usage

The old interface used the full area of the display hardware. This has the disadvantage that there is no additional space to put other interface elements without hiding GPR information. The old interface shows significantly more information to the operator than is strictly necessary for the classification of the signal which triggered the detection.

The new interface uses a square display of the same height as the old interface. The square layout is made possible by the relative width of the track of the sensor compared to its turning radius. Since the track points down on the display and curves at a maximum rate corresponding to about 3.5 meters of displacement at the bottom of the display, the left and right area of a fullscreen map would never have data. Removing these regions and making the map square frees up 40% of the horizontal space on the interface, which could be used for other purposes such as the status information and alerts. Keeping the status information constantly available would help decrease the number of nested layers of elements in the interface.

3.6 Summary

This chapter presented a human factors analysis of the Gopher system. Based on that analysis it presented a model of the human operator's behavior and interaction with the system. From this model, a set of requirements were distilled and documented. A set of new interface features were presented and compared to the old interface design. This interface will be tested in an experiment described in Chapter 4 and the results will be presented in Chapter 5.

Chapter 4

Usability Analysis Experiment

4.1 Introduction

Chapter 3 proposed a new design for the GPR display component of the Gopher system. An experiment was conducted that compared the usability of the new interface to the old interface. This comparison was made by showing test subjects the same scenarios of GPR data with two different interfaces and asking them to classify the data as depicting an object or a false alarm. This chapter presents the design of the experiment and details about how it was conducted. The results of this experiment are presented in Chapter 5.

4.2 Participants

Thirty-four subjects from MIT Lincoln Laboratory were recruited over email and through word of mouth. Three additional subjects served as pilot testers. Three subjects were considered experts due to prior affiliation with the project. Two subjects were disqualified after the fact due to colorblindness. Removing the experts and disqualified subjects yielded a pool of thirty experimental subjects. The questionnaires used to obtain this information are found in Appendix A.

Out of thirty subjects, twenty-one were male and nine female. Twenty-two subjects self-identified as research staff members. Six subjects self-identified as non-research staff. Two subjects self-identified as students. Two subjects self-reported military

service. Fourteen self-reported having worked with or studied radar systems before. Subjects ranged in age from twenty-two to seventy-two (μ: 41 years, σ: 13 years). All subjects were fluent in written and spoken English with self-reported corrected vision 20/25 or better.

Questionnaire responses indicated that most subjects use a touchscreen device daily, do not play video games frequently, and described themselves as conservative decision makers. See Tables 4.1, 4.2, and 4.3 for tabulations of subject responses.

Never	A few times ever	A few times a year	Daily
2 (7%)	1 (3%)	6 (20%)	21 (70%)

Table 4.1: Subject self-reported Touchscreen Use

Rarely	Once a month	Weekly	A few times a week	Most days
21 (70%)	2 (7%)	3 (10%)	1 (3%)	3 (10%)

Table 4.2: Subject self-reported Video Gaming Frequency

Very conservative	Conservative	Neutral	Risky	Very risky
0 (0%)	19 (63%)	9 (30%)	2 (7%)	0 (0%)

Table 4.3: Subject Self-reported Decision-making Conservativeness

Additional information and the raw demographic data of the subjects is available in Appendix D.

4.3 Testbed

The Gopher system uses a commercial flatscreen display with resistive touchscreen capabilities. For convenience and cost savings, this experiment was conducted using the Apple iPad® as a testing platform. According to manufacturer specifications, these two displays have the same pixel density, 52 pixels per centimeter. While there are differences between these two hardware setups which may impact other areas of human factors research, they were not relevant to the key questions this experiment hoped to explore.

Apple Keynote® was used to display training information to subjects on the iPad®, and a custom app was written to present the experiment to subjects.

4.4 Experiment Design

A fully-crossed experimental design was chosen to enable comparisons between subjects and between interfaces within subjects. Subjects were given 102 scenarios to classify, once on each interface, for a total of 204 scenarios per subject. The order of the scenarios was randomized for every subject and between the two interfaces. Subjects were shown all 102 scenarios on one of the two interfaces first and then the other interface. 'First' and 'Second' will be used to describe the order in which the interfaces were shown to a particular subject. 'Old' and 'new' will be used to describe the two interfaces regardless of the order they were shown. Subjects were counterbalanced to minimize gender and treatment order impact on the results.

Each subject was put through the same procedure detailed in Appendix A, summarized as follows:

1. Demographic survey
2. Self-guided training slides for the first treatment.
3. Practice session for the first treatment.
4. First treatment trials
5. First post-trial survey
6. Self-guided training slides for the second treatment.
7. Practice session for the second treatment.
8. Second treatment trials
9. Second post-trial survey

Subjects met the experimenter in one of several small rooms at MIT Lincoln Lab depending on room availability. Almost all of the subjects took the experiment

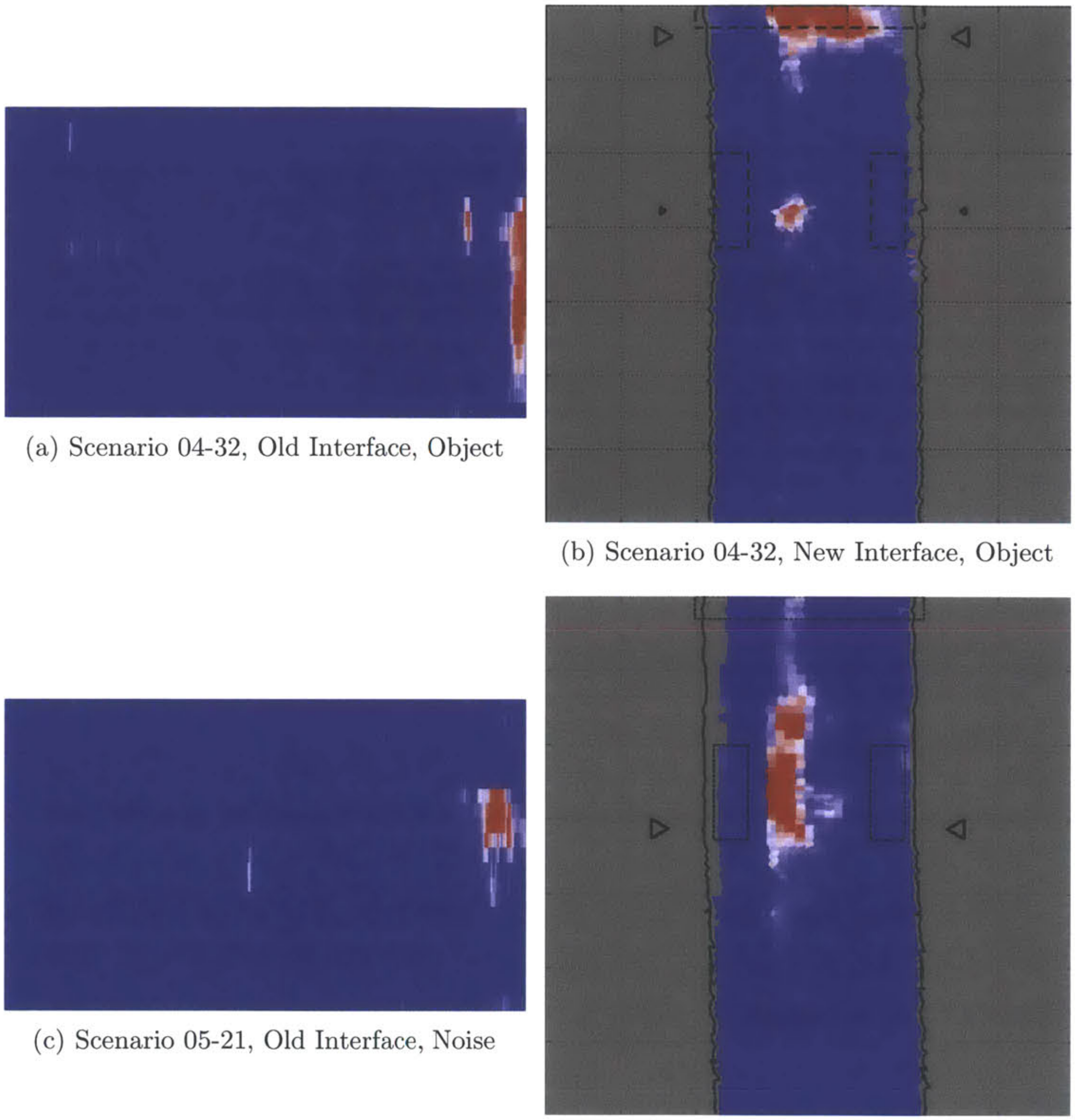
individually, though for scheduling purposes pairs were permitted to be in the same room at staggered times. No subjects were interrupted during the course of the experiment. All but two subjects finished within the intended hour timeframe. At the beginning of the experiment, consent was obtained and subjects were presented with the brief demographic survey.

Subjects were presented with a slide deck on the iPad® to explain the goals of the experiment and each interface. See Appendix B for a copy of the slides. Subjects navigated through the slide deck at their own pace and could spend as much time as they wished looking at the materials. The last ten slides in the training materials were scenarios with the correct classification labeled. Subjects could ask questions at any time except during the trial phases of the experiment. Subjects were encouraged to ask questions at the end of each phase. Subjects were informed that their performance would only be assessed by the correctness of their answers, and not by the time they took to complete the exercise. Subjects were also informed of a $100 prize for the best performance.

Following viewing the slide show, each subject was given a brief practice opportunity in which ten training scenarios were shown on the interface exactly as if it were the trials. The ten training scenarios were the same scenarios as the last ten slides of the training slide materials, though subjects were not told that in advance. After submitting each answer in the training sequence, subjects were shown the correct answer. At the end of the ten scenarios, a summary screen displayed the results and the experimenter talked through the incorrect answers with the subject and answered any questions. If the subject incorrectly answered three or more scenarios, a remedial training protocol was then given which presented an additional six scenarios. If the remedial scenarios were shown, the experimenter reviewed their performance again at the end of the six slides. Three subjects required the remedial training protocol for one of the two interfaces, and one subject required the remedial training protocol for both interfaces.

At the end of the training procedure, the subjects were given the full 102 scenarios to classify, which included the scenarios from the training, randomly interspersed. See

Appendix C for the full set of scenarios. These scenario interfaces were generated using the process explained in Appendix E.2. Four examples are shown here in Figure 4-1 to illustrate what the user was shown.



(a) Scenario 04-32, Old Interface, Object

(b) Scenario 04-32, New Interface, Object

(c) Scenario 05-21, Old Interface, Noise

(d) Scenario 04-32, New Interface, Noise

Figure 4-1: Two Example Scenarios on Both Interfaces

Every ten scenarios, a slide was shown which suggested the subject could take a break. Subjects were not shown the correct answers for each scenario as they were

during training. Each scenario was presented and subjects could press one of two buttons indicating the answer they believed to be correct. Subjects were not able to change their answer after entering it the first time. See Figure 4-2 for an example of the interface during this process.

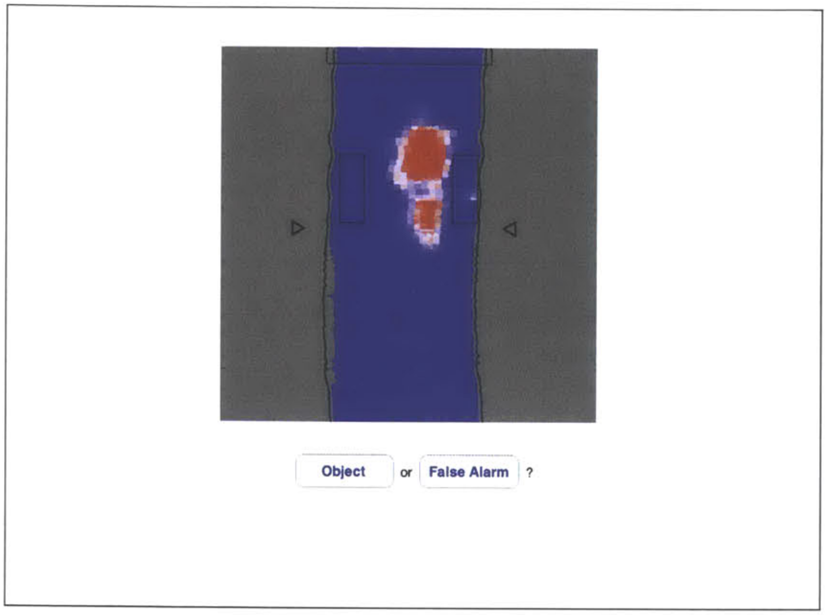


Figure 4-2: iPad Interface Showing the New Display: Asking for a Decision

After each answer was collected, a slider was displayed and the subject was asked to rate his or her confidence in his or her given answer on a seven value scale. See Figure 4-3 for an example.

After each set of trials, subjects completed a brief questionnaire. The total time for the experiment was not precisely controlled since subjects could set their own pace. Subjects were told to expect the exercise would take about an hour, and almost all the subjects finished within that time. Subjects were compensated for their time in the form of items available from the MIT and MIT Lincoln Laboratories gift stores--mugs, shirts, and umbrellas--beverages and pastries valued at $15. To motivate subjects, a grand prize for the highest performer was a $100 BestBuy® gift certificate.

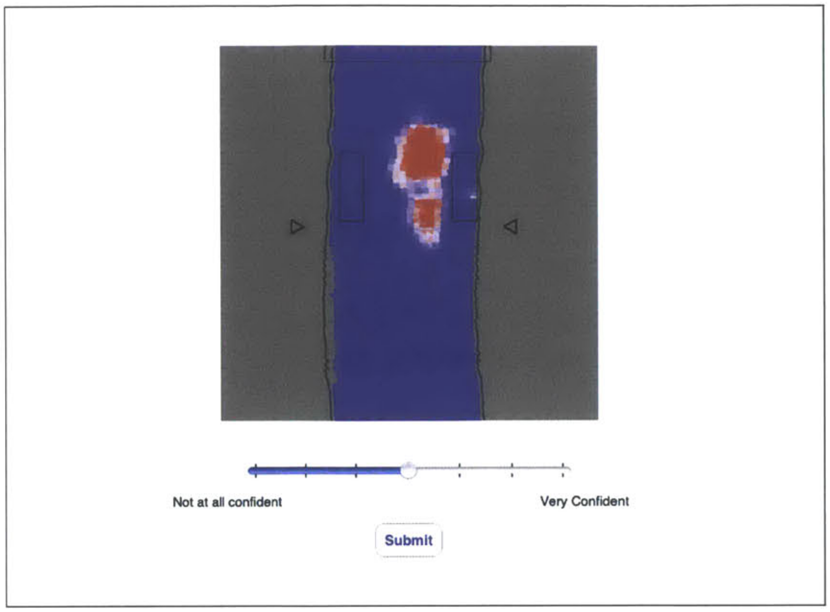


Figure 4-3: iPad Interface Showing the Old Display: Asking for Confidence

4.5 Output

Each subject produced four sources of data. Three sources were obtained from surveys covering demographic data, prior experiences, opinions about each interface, and self-assessments of performance. The surveys are available in Appendix A. The fourth source was the log file of the subject's training and trial interface interactions.

The log files of the training and trials contained timestamped events whenever:

- An example was shown to the user
- An answer was entered by the user
- A confidence level was entered by the user
- A break started (preprogrammed, every ten examples)

4.6 Summary

This chapter presented the experiment designed to test the new interface described in Chapter 3. The experiment used a fully-crossed design to test all subjects with

both interfaces for between subjects and within subjects comparisons. It investigate the effects of the interface changes and how demographics affect subject performance. The experiment was conducted with a pool of thirty volunteer subjects using implementations of both interfaces on an iPad. The results of this experiment are presented in Chapter 5.

Chapter 5

Results and Discussion

5.1 Introduction

This chapter presents the results of the experiment motivated by Chapter 3 and detailed in Chapter 4. A summary of the raw data used to calculate statistics and render figures is included in Appendix D. Additional commentary on the implications of these results and future steps are found in Chapter 6.

5.2 Interface Comparison

5.2.1 Correctness

Subjects correctly categorized more scenarios using the new interface (two-tailed two-sample t-test, t=2.658, p=0.013). The old interface had a mean correct percentage of 66.6% and the new interface had a mean correct percentage of 69.7%, a difference of 3.1% of the total scenarios, or 4.7% improvement. These results are shown in Figure 5-1.

5.2.2 Misses

Subjects had significantly fewer misses on the new interface (two-tailed Wilcoxon signed-rank test, z=3.766, p=0.000). The old interface had a mean miss percentage of 18.5%, and the new interface had a mean miss percentage of 13.12, a difference of 5.38% of the total scenarios, or a 29.1% improvement. See Figure 5-2;

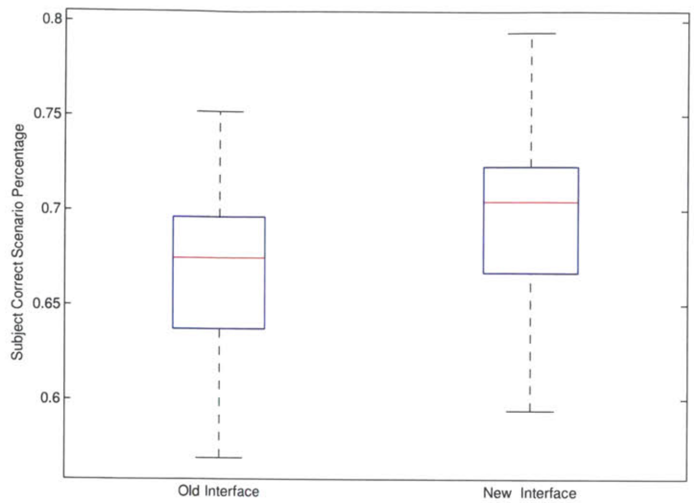


Figure 5-1: Correct Ratio for Subjects on Both Interfaces

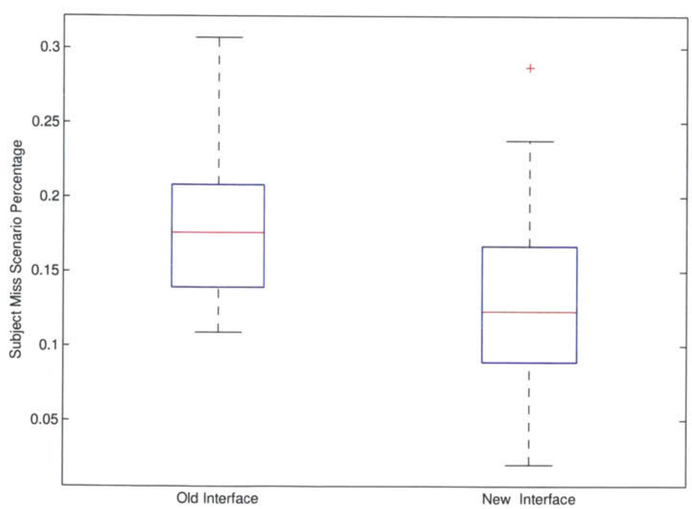


Figure 5-2: Miss Ratio for Subjects on Both Interfaces

5.2.3 Confidence

Insufficient evidence was found to suggest subjects were more confident in their answers on either interface (two-tailed two-sample t-test, t=1.920, p=0.065). However a significant linear correlation was observed between the average confidence on the interfaces (p = 0.000, $r^2 = 0.726$). Spearman's rank correlation coefficient indicates these two metrics are highly related (p=0.000, $\rho = 0.866$).

When confidence and correctness were examined together, a significant difference between the two interfaces was observed. A metric was calculated to assess the appropriateness of each subject's confidence in their answers. This metric was calculated as follows:

$x_i \equiv 1$ if the subject answered scenario i correctly, or -1 otherwise
$c_i \equiv$ the subject's confidence $[1-7]$ in their answer for scenario i

$$\text{confidence-correctness} = \frac{\Sigma x_i c_i}{\Sigma c_i}$$

A subject with a higher confidence-correctness score attributed higher confidence scores to the scenarios he or she answered correctly. Scaling by the sum of the subject's confidence scores attempts to normalize subjects with different internal models of confidence.

Subjects had higher confidence correctness scores on the new interface (two-tailed two-sample t-test, t=3.745, p=0.001). Subjects had a mean confidence-correctness score on the old interface of 0.377, compared with 0.462 on the new interface.

5.2.4 Estimated Performance

Insufficient evidence was found to suggest that subjects estimated their percentage of correct answers differently between the two interfaces (two-tailed Wilcoxon signed-rank test, z=1.489, p=0.136). This is interesting because subjects had higher confidence-correctness scores on the new interface. Since there was an improvement in confidence--correctness and little change in average confidence or self-estimated performance

estimates, then it suggests that subjects were simply more accurately assessing their correctness.

5.2.5 Learning Bias

Since this experiment compared performance within subjects in addition to between subjects, learning bias was a potential confounding factor. Subjects were either presented with the old interface first or the new interface first and then were shown the other interface. Therefore it is reasonable to expect that seeing the 'first' interface might alter the experience of using the 'second' interface, introducing effects that were not related to the experience of using either interface in isolation.

Evidence of learning bias was found in the data, but it did not impact the results in a meaningful way. The only form of learning bias found was a small efficiency increase on the second interface; subjects had a higher mean time per scenario on the first interface regardless of which interface type was shown first (two-tailed Wilcoxon signed-rank test, $z=4.001$, $p=0.000$). The mean average time (the mean value of each subject's mean time to categorize a scenario) for subjects on their first interface was 4.8 seconds. The mean average time for the second interface was 3.8 seconds. This is reasonably explained by the hypothesis that subjects became more adept at using the experiment interface and fell into a faster rhythm.

5.3 Demographic Factors

The data from this experiment permits analysis of various metrics with respect to demographic differences. Although this study was not specifically designed to test for demographic factors, the statistical assumptions for such analyses are met. In order to make this analysis more meaningful, subject responses were grouped into binary categories based on the distribution of the responses. Table 5.1 shows the results of independent sample t-tests and Mann-Whitney U tests (depending on metric normality) for the demographic variables. The groupings used were:

- Job: Subjects were divided into technical staff and others. This grouping was chosen because there was reason to believe that highly-educated career

researchers may behave differently than career administrators due to education and past experience with plots.

- Gaming frequency: Subjects were divided into low-gamers (less frequently than once-a-month) and high-gamers (once-a-month and more frequently).

- Conservativeness: Subjects were divided into very conservative through conservative and neutral through risky.

- Touchscreen usage: Subjects were divided into frequent (daily) users of touchscreens and all other frequencies.

The most notable pattern from these results is that women performed better than men in many respects. For the new interface, women answered more scenarios correctly, missed fewer objects, had a higher confidence correctness score, and had a higher average confidence. For the second interface they used, women predicted that a higher percentage of their answers were correct, and had a higher average confidence.

Technical staff (researchers and engineers) demonstrated more correct answers compared to other users on the old interface. They also had higher confidence-correctness scores compared to other users on the old interface. These results were not predicted, but are unsurprising as familiarity with nonintuitive plot formats is a skill one would expect from engineers and scientists.

Subjects that self-described themselves as more conservative than other subjects missed fewer objects in the first interface they interacted with and had higher average confidence in the second interface they used. This suggests that personality traits affected the way that users interacted with a challenging situation and how they continued to interact with it after a prolonged period of time.

Subjects with more frequent video gaming habits were faster on the first interface they interacted with compared to users with less frequent video gaming habits. This suggests that experience with games causes users to speed up their interaction with a new system, though the difference was not detectable in the second interface (regardless of which interface was first or second).

Subjects that indicated they had not previously worked on radar systems missed fewer objects on the new interface. This is mildly surprising as it suggests that there is something counterintuitive about the new interface that might cause prior experience to be detrimental to the use of this interface design.

Subjects that more frequently use touchscreen devices such as tablets or smartphones missed fewer scenarios on the second interface and had higher confidence-correctness scores on the second interface. It is unsurprising that familiarity with a touchscreen system might cause users to behave differently, though there is no obvious connection between touchscreen use and these metrics.

Table 5.1: Metrics and Demographic Factors

Metric	Interface	2-t t/M-W U	p	Value	Value	Demographics and difference of the means between groups
Correct Ratio	Old	t=2.19	.037	.638	.667	Technical staff had 2.9% more correct scenarios.
	New	t=3.20	.003	.681	.735	Women had 5.4% more correct scenarios.
Miss Ratio	New	U=164	.031	.106	.160	Subjects without radar experience missed 5.4% fewer scenarios.
	New	U=44.5	.022	.133	.167	Women missed 3.4% fewer scenarios.
	Second	U=35.5	.006	.131	.216	Subjects with more touchscreen experience missed 8.5% fewer scenarios.
Est. Correct	Second	U=48.0	.036	61.5	72.0	Women estimated 10.% more correct scenarios.
Conf.-Correct.	Old	t=2.60	.015	.310	.402	Technical staff had a .092 higher confidence-correctness score.
	New	t=2.40	.023	.434	.528	Women had a .094 higher confidence-correctness score.
	Second	t=2.79	.009	.352	.469	Subjects reporting more frequent touchscreen use had a .12 higher confidence-correctness score.
Mean Confidence	New	t=2.65	.013	4.46	5.13	Women had a .67 higher average confidence.
	Second	t=2.37	.025	4.38	5.00	Women had a .62 higher average confidence.
	Second	t=2.12	.043	4.39	4.72	Conservative subjects had a .33 higher average confidence.
Mean Time (s)	First	U=45.0	.025	3.71	5.79	Subjects with more gaming experience had a 2.1 second lower mean time.

5.4 ROC Analysis

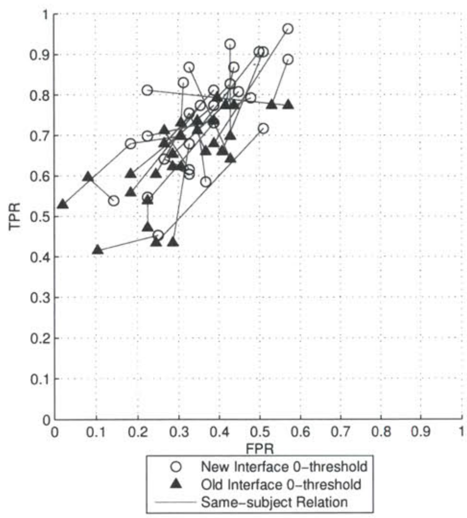
Considering only the binary decision made by subjects and ignoring their confidence scores, Figure 5-3 shows subjects TPR and FPR values for both interfaces, linked together with a line. The details of these calculations and metrics can be found in Section 2.2.1 Since subjects reported confidence scores in addition to their binary answers, it is possible to generate an ROC curve treating each subject's confidence value as a stimulus. This was accomplished by adapting every subject's binary answer and confidence score into a sign-extended confidence score on the interval $[-7, 7]$ and varying the threshold for a decision. These curves are also depicted in Figure 5-3. In addition this figure also shows the ROC points generated with a threshold of zero, corresponding to the binary decision given by subjects ignoring their confidence.

Gopher's developers favored high TPR values from a policy standpoint. This is because the existing policy for operators was "investigate everything". This policy is the same as declaring all detections positive. Therefore the point of comparison for all potential variants on the system is (100% TPR, 100% FPR). Consequently, any substantial reduction in FPR with an acceptable loss in TPR is considered a success from the Gopher project's policy standpoint.

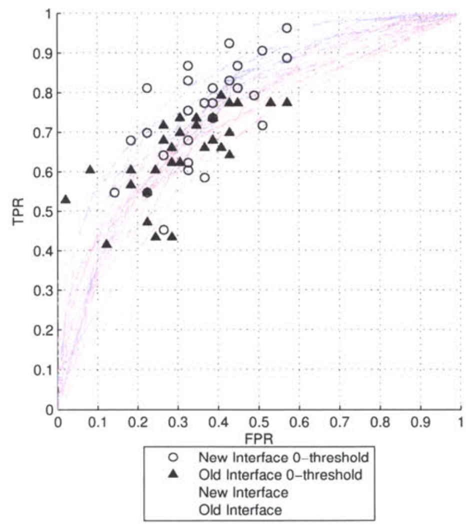
These ROC curves visually illustrate trends shown by statistics mentioned earlier. The new interface is generally superior as shown by more of the subject ROC curves being tighter against the top and left axes. This indicates that for a particular FPR value, new interface achieved a higher TPR. Likewise, for a particular TPR value, the new interface achieved a lower FPR. The new interface is generally more prone to false positives than false negatives indicated by the 0-threshold points being further up and to the right than the 0-threshold points for the old interface.

These curves also demonstrate the general category of achievable classifier performance. Notably, points in the realm of 95% TPR at approximately 50% FPR on the new interface are possible. While these numbers may appear deficient to some audiences, when compared to a policy of "investigate all signals" they represent a halving of the FPR with negligible loss to TPR. These results provide a rough

estimate of the performance of this system.



(a) Subject ROC 0-threshold values linked between interfaces



(b) Subject ROC 0-threshold values and ROC curves

Figure 5-3: Subject ROC Values and Curves

5.5 Collaboration and Collective Intelligence

In addition to evaluating individual subjects, this data set enabled exploration of the collective performance of subjects. There are many approaches that can be used to boost the input of users into a single decision. One approach was to form committees of various sizes and enable each subject to vote. Using a confidence voting mechanism, committee ROC curves were obtained using randomly selected independent committees of various sizes. These committee ROC curves and original subject ROC curves are shown in Figure 5-4.

These committee curves illustrate several interesting points. Generally, for large committee sizes, a committee's performance is asymptotically equivalent to the best performers in the experiment at TPR values of 90% and higher. The precise performance numbers vary every calculation due to randomization of the committee

memberships. Points such as (100% TPR, 63% FPR) and ($\geq$98% TPR, $\leq$50% FPR) ($\geq$90% TPR, $\leq$40%) are typical. As the committee sizes drop, greater variance occurs between committees of the same size.

Since individual performance affects the performance of a committee, committees with higher performing members will perform better than committees with less capable members. Figure 5-5 shows committees of various sizes formulated from the top and bottom performers. For ranking purposes, performers were ordered by the total area under the convex hull of their ROC curve to determine their membership in these committees.

This comparison demonstrates that the new interface is superior to the old interface in three manners.

- The best performers and committees of best performers on the new interface are superior to their counterparts on the old interface.

- The worst performers and committees composed of the worst performers on the new interface are superior to their counterparts on the old interface.

- For large committee sizes (10 and 15), the worst performers on the new interface performed better than the best performers on the old interface at TPR values of 90% and higher.

The most interesting result of the committee analysis is the performance of the best and worst 15-member committees. As shown by the 15-member ROC curves, the top and bottom 50% of subjects collectively performed asymptotically well at TPR values greater than 90%. This is an astounding result, as it suggests that the benefits of collective intelligence significantly outweigh the magnitude of individual performance variation as seen by comparison to the lowest performing 50% of subjects. Furthermore, the performance of both 15-committees is better or equivalent to those of the top performers for the respective interfaces. Figure 5-6 demonstrates this comparison explicitly.

5.6 Expert Users

Two subjects (beyond the thirty subjects previously analyzed) were expert users of the system. These experts were heavily involved in the design and testing of the Gopher system. Figure 5-6 shows the ROC performance of these experts in comparison to several previously mentioned committees and the top performers. It is notable that these experts performed less well than the committees of best and worst 15, in addition to performing less well than the entire 30-committee. Individually, the experts were ranked 5 and 20 out of 32 on the old interface, and they were ranked 10 and 11 out of 32 on the new interface. This indicates that the experts were potentially better than many subjects, but still outperformed by many others. This further demonstrates the value of the committee concept by showing that the collective ability of the lowest performing 50% of subjects performed better than the experts and creators of the system in the high-TPR region.

5.7 Scenario Analysis

Different scenarios were the most challenging to users on each interface. There was only a moderate correlation between the percentage of users correctly answering the same scenarios on the old interface and the new interface (linear fit, $r^2 = 0.483$). When the top five scenarios most likely to cause a hit on each interface were compared, two scenarios appeared in both lists. The top five scenarios for misses, false positive, and correct rejections shared no common scenarios. This demonstrates that the frequent offenders for both interfaces were different rather than similar.

Figures 5-7, 5-8, 5-9, and 5-10 show the top five most frequently classified scenarios in each of the four possible binary decision outcome categories, hits, correct rejections, false positives, and misses. These scenarios clearly illustrate several phenomena:

- Figure 5-7: The easiest object signals to correctly identify (hit) were single, isolated blobs with a moderate sized peak and smooth gradient surrounded by little noise.

- Figure 5-8: The easiest noise signals to correctly identify (correct rejection) were

scenarios in which the interface showed no gradient around peak values and placed small regions of high values (red speckles) next to large areas of low value (blue) or scenarios in which the vehicle was significantly off track. In three of these top five scenarios, the vehicle was turning or having difficulty matching the baseline path.

- Figure 5-9: The noise signals problematic for subjects (false positive) were small signals, sufficiently large to cause users to identify them as objects, often in clusters.

- Figure 5-10: The object signals most problematic for subjects (false negative) were objects that produced small visual anomalies similar to nearby clutter and noise. In several cases there were also additional high values further along the track from the detection event. It appears that these later signals masked the presence of the actual target.

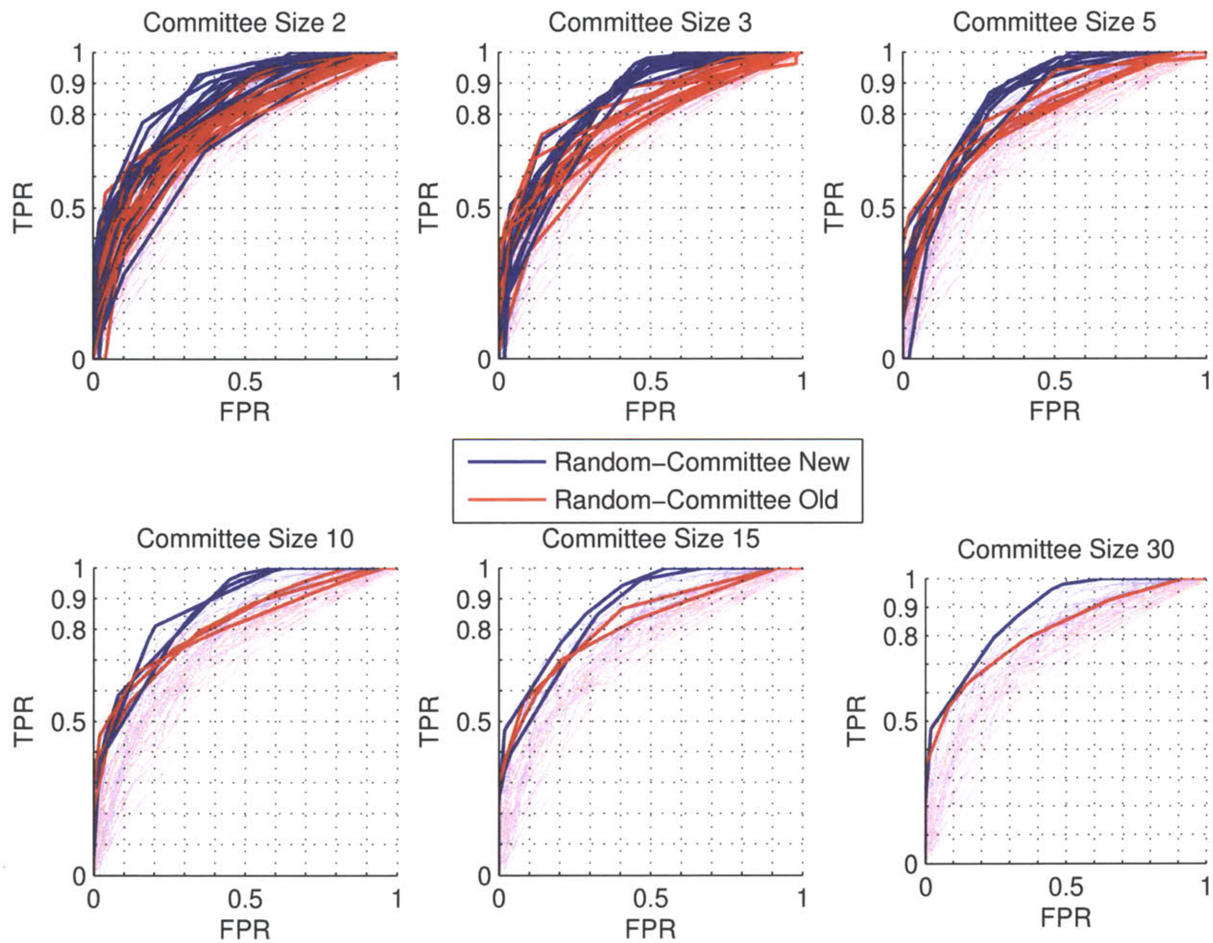


Figure 5-4: Committee ROC Curves for Sizes 2, 3, 5, 10, 15, 30

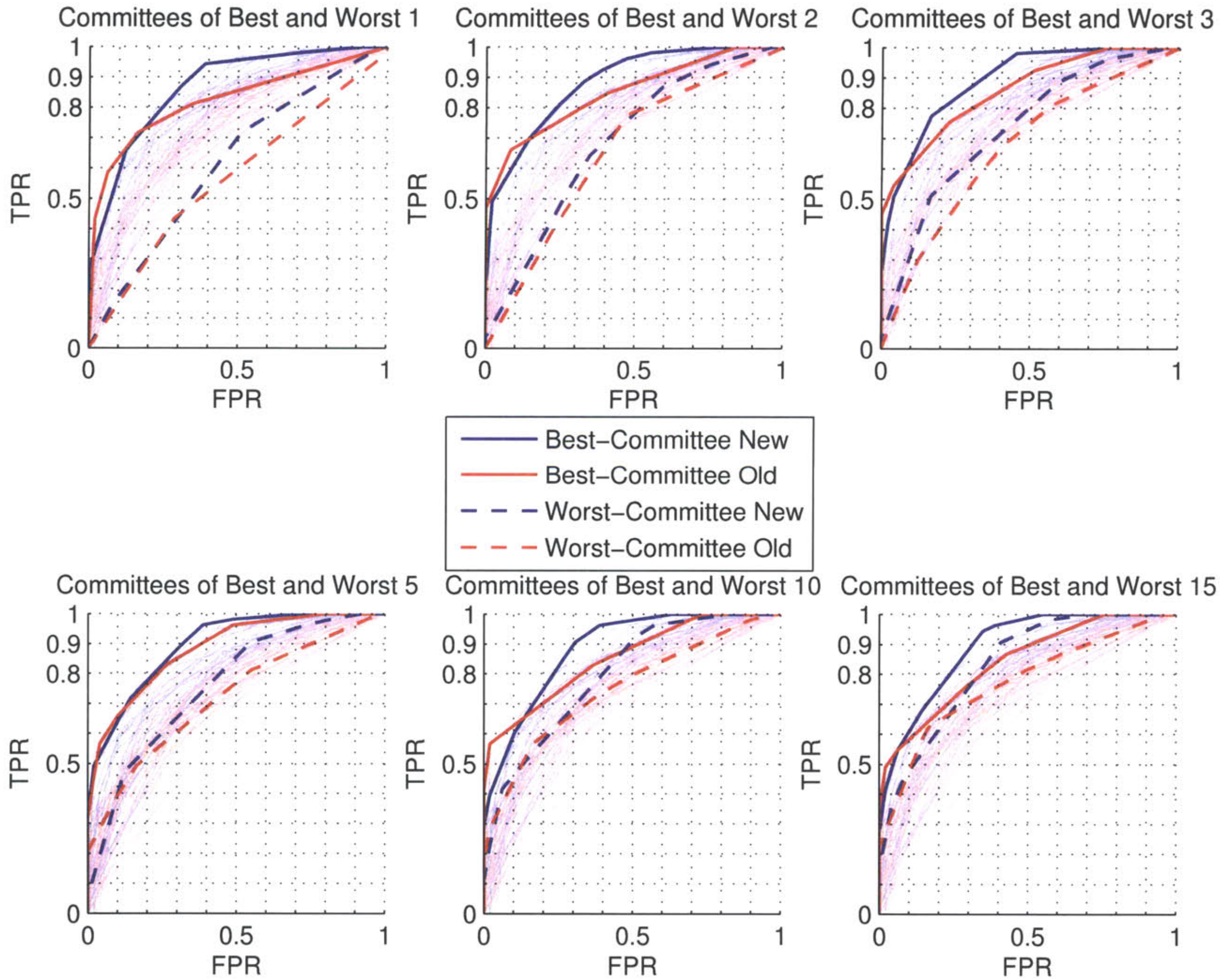
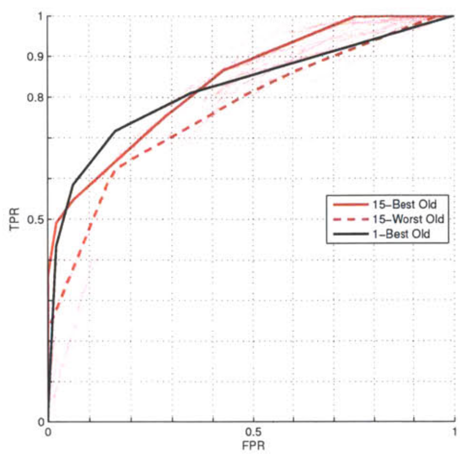
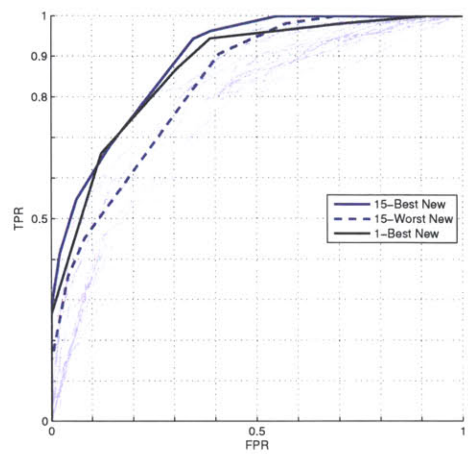


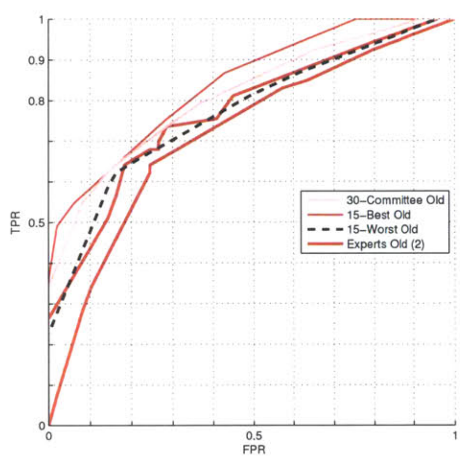
Figure 5-5: Best and Worst Committee ROC Curves for Sizes 1, 2, 3, 5, 10, 15



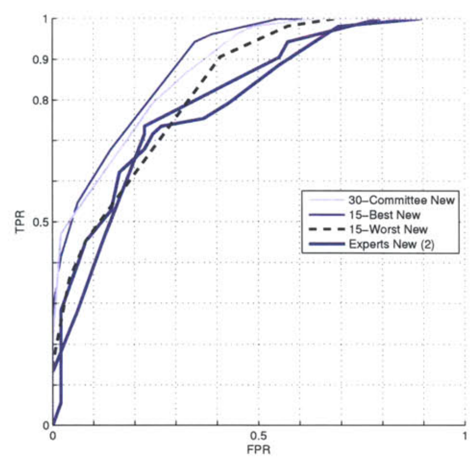
(a) ROC curves of committees for the old interface



(b) ROC curves of committees for the new interface

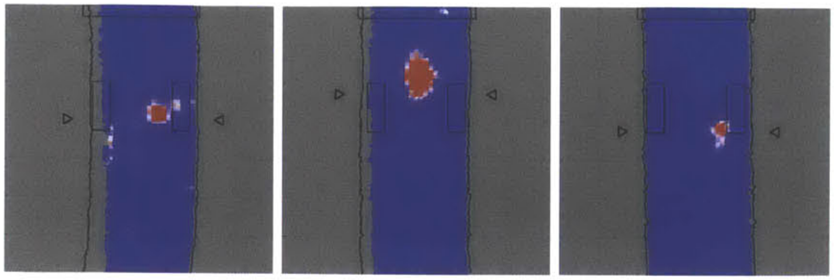


(c) ROC curves of committees and experts for the old interface

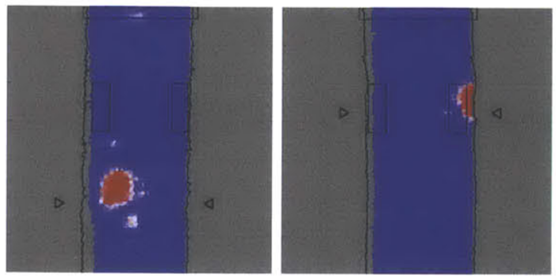


(d) ROC curves of committees and experts for the new interface

Figure 5-6: Selected ROC Curves and Experts

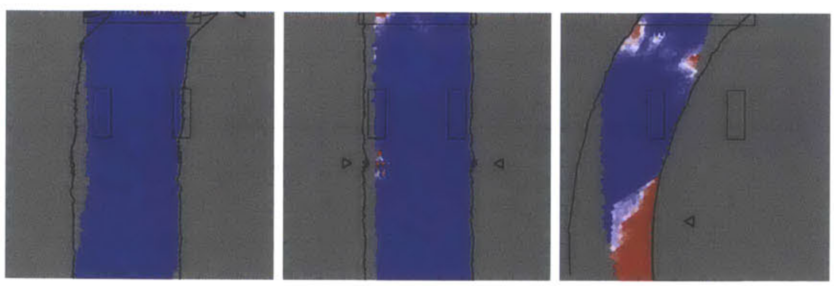


(a) Scenario 22-14 (b) Scenario 18-05 (c) Scenario 14-17

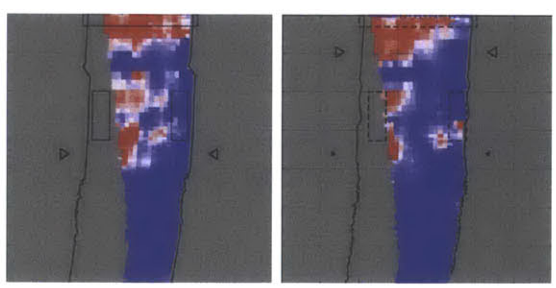


(d) Scenario 01-11 (e) Scenario 01-07

Figure 5-7: Top Five Hit Scenarios on Interface B

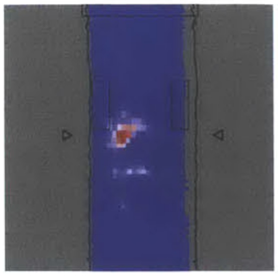


(a) Scenario 12-40 (b) Scenario 08-37 (c) Scenario 01-13

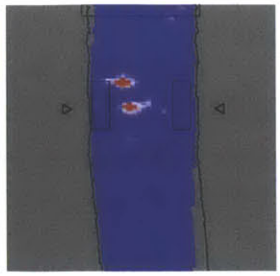


(d) Scenario 16-43 (e) Scenario 07-47

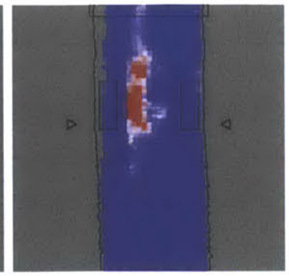
Figure 5-8: Top Five Correct Rejection Scenarios on Interface B



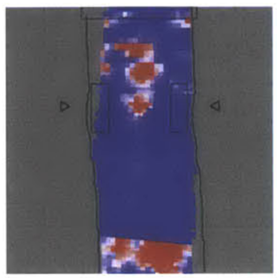
(a) Scenario 14-24



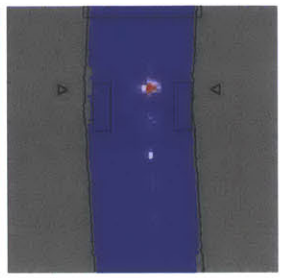
(b) Scenario 09-17



(c) Scenario 05-21

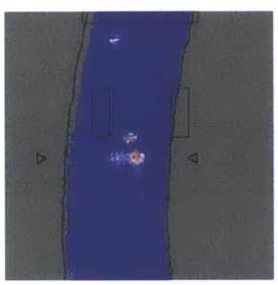


(d) Scenario 03-17

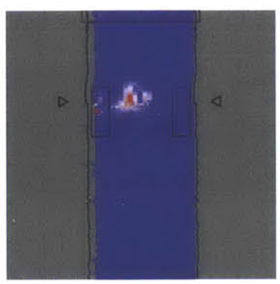


(e) Scenario 11-02

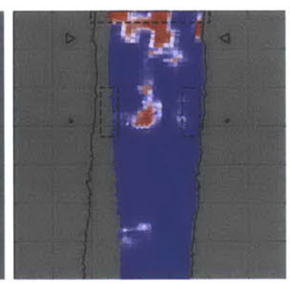
Figure 5-9: Top Five False Positive Scenarios on Interface B



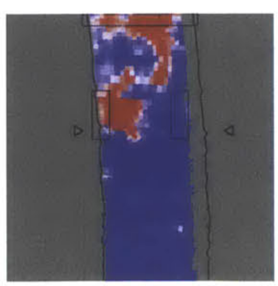
(a) Scenario 16-28



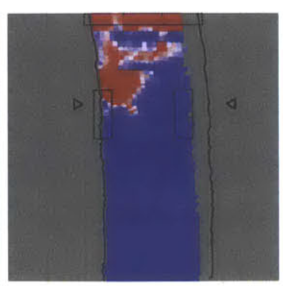
(b) Scenario 10-34



(c) Scenario 04-22



(d) Scenario 03-15



(e) Scenario 09-15

Figure 5-10: Top Five Miss Scenarios on Interface B

5.8 Subject Feedback

Subjects overwhelmingly preferred the new interface. When asked on a five-value Likert scale to assess their preference between the interfaces, subjects responded as follows:

Strongly prefer old	Prefer old	No pref.	Prefer new	Strongly prefer new
1 (3.3%)	2 (6.7%)	2(6.7%)	8 (26.7%)	17 (56.7%)

Table 5.2: Subject Interface Preference

A total of twenty-five out of thirty subjects indicated they preferred the new interface compared to the old interface.

Subjects were asked several qualitative questions about the interfaces after each trial. The questions can be found in Appendix A. The qualitative feedback from subjects agreed with the analysis of the most frequently categorized scenarios presented in the previous section. Subjects felt that single isolated signals were the easiest to identify. Multiple signals and complex signal shapes were considered harder to classify. Subjects reported that seeing massive signals or multiple tiny signals was a clear sign that the scenario was a false alarm. Signals which bordered the edge of the data region were regarded as harder to classify.

Subjects appreciated the triangles which identified the values triggering the automated detection system. In contrast, subjects complained that the old interface gave them no indication which areas of the plot were the most relevant. Subjects commented that the aspect ratio of the pixels in the old display complicated their analysis compared to the nearest-neighbor pixels of the new display. Subjects had varied preferences for the direction of the display, though most preferred the track-up layout of the new interface rather than the track-right orientation of the old display. At least two subjects did rotate the iPad® by a full 90 degrees for several scenarios during the exercise, though frequently that was associated with their posture in their chair rather than purely the interface direction. One subject that rotated the display

during his new interface trial was an expert and therefore more familiar with the track-right display design.

The qualitative answers provided an important window into human intelligence. One of the traditional approaches to artificial intelligence research is to mimic human intelligence. In order to mimic human behavior, an understanding of human decision-making mental models is required. One unstated goal of the survey questions was to investigate the language and features used by human subjects to analyze these signals. While subjects were given training slides that showed scenarios before each interface exercise, an absolutely minimal mental model was given to them about what aspects of the signals were most relevant.

From all thirty-seven subject response surveys, some of the most common words in responses included: "clear" (eight subjects), "smeared" (four subjects), and "blob" (fourteen subjects). These words were used to describe the most challenging and easy scenarios to classify in addition to the relationship between peaks and the surrounding data for objects and false alarms.

5.9 Summary

This chapter presented the results of the experiment described in Chapter 4.

The individual performance of subjects on both interfaces was examined for differences. Subjects were found to perform better on the new interface design. In particular, subjects made more correct decisions on the new interface, and they also missed fewer objects. Subjects were also found to have more accurately assess their own performance based on their self-reported confidence scores.

Subjectively, subjects overwhelmingly preferred the new interface. Subjects appreciated several of the new features, most notably the alerting indicators that pointed out the relevant region of the display.

Various demographic factors were examined for their correlation with subject performance. Technical staff faired better with the old interface. Women performed better than men in several metrics on the new interface including higher confidence, higher confidence-correctness. Subjects that played video games more frequently were

faster to complete the first set of scenarios.

The collective intelligence of the subjects was examined through the use of voting committees. Committees of various compositions including randomized, best, and worst ranked subjects were were examined. Committees performed better on the new interface than the old interface, particularly at TPR $\geq 90\%$. The committees were also compared against top performers and experts. Large committees were shown to perform better than the top performers and expert users at high TPR values.

The scenarios that were most likely to cause subjects to classify them as hits, correct rejections, false positives, and misses were examined. These scenarios showed several trends, which matched the subjective feedback of subjects. Single isolated signals were the easiest to identify as objects, while massive signals were clearly false positives. Multiple signals and faint signals were the easiest to misclassify.

Chapter 6

Conclusions

6.1 Research Objective Findings

The research objectives of this work were:

- *Objective 1*: Determine the functional and information requirements for a vehicle-based CD GPR signal analysis tool.
- *Objective 2*: Design an improved interface which supports human analysis of GPR signals.
- *Objective 3*: Evaluate the effectiveness of the interface design through experimentation and the use of credible metrics.

The results of Objective 1 and 2 were presented in Chapter 3. The key interface requirement identified for Gopher consisted of ensuring that the display provided the necessary information to support the operator's detection capabilities and reporting of detections. In response to these requirements, the focus of the research was directed towards the interaction of user with the signal classification process using radar imagery. A new interface was designed that leveraged spatial display of radar data to both assist in signal classification and the communication of signal location to others. The new interface design was implemented and then analyzed in an experiment with human subjects described in Chapter 4.

The results of Objective 3 were presented in Chapter 5. In summary, the new proposed interface was objectively better than the previous interface in some, but not all performance metrics. The most important performance increase was a 29% decrease in missed objects on the new interface. Other advantages included a better overall correctness. Subjectively, subjects preferred the new interface by an overwhelming majority. This combination of qualitative and quantitative metrics shows that the novel spatial display has significant benefits compared to the temporal display.

In addition to these objectives, an unexpected result was found in the collective performance of subjects. Using data from the experimental work taken from subjects in isolation, committees were shown to substantially augment the performance of the system compared to individual operators. The fact that committee performance exceeds that of top performers suggests that in the future, resources may be better spent implementing a system designed to leverage multiple user's abilities rather than training specific individuals to operate in isolation.

6.2 Future Work

Many areas of this thesis sparked questions and further avenues of exploration. The most interesting areas for additional research are summarized here.

6.2.1 Color Mapping

The displays used in this experiment all had the same mapping of radar values to color. That mapping is linear between two arbitrary points, rounding all values past the color range at the extremes. Further research should explore different mappings of radar values to color. A mapping which gives greater color space to the lowest radar values would help in discriminating between faint targets and faint false alarms, while that same mapping would fail to visually distinguish a medium-strength false-alarm from a high-strength target. Spreading the same color range over a wider range of radar values would help distinguish high-strength signals from medium-strength signals, but it would reduce the resolution between neighboring color values. Potential mapping ideas include nonlinear mappings or discontinuous mappings. Various color

combinations could also be tried. This research would benefit from real lighting conditions inside the vehicle rather than an artificial test environment.

6.2.2 Display Zoom

The new display used a fixed seven-meter by seven-meter map. This size was chosen because it was as large as feasibly possible without causing the tracks of the sensor to intersect the side of the display. It is possibly that alternative sizes, or variable zoom, might be useful in some applications. Additional studies or walkthroughs with various operators could identify an optimal zoom to balance the need for detail with the need to understand the biggest picture possible with a given screen size.

6.2.3 Correct Answer Ratio

This research trained and tested users with an equal ratio of targets to noise over the course of the experiment. Therefore users were not statistically biased to expect one result more frequently than the other. If users are shown a biased mix of signals to classify, it is very possible that their ability to correctly classify scenarios will diminish due to alarm fatigue or complacency. Further research should examine what happens to the performance characteristics of operators when they are shown a biased mix of mostly-noise or mostly-objects.

6.2.4 Target Variety

This research used data from a particular set of buried objects. The training slides for subjects included this same set of objects. If this research was applied to larger objects, such as tunnel detection, additional tools and interface elements would be needed to differentiate signals from noise. In this data set, one easy way to identify some types of false alarms was purely based on signal size. Since there were no objects of similar size to the vehicle in the training scenarios, large signals were easily classified as false positives. Should the system encounter a real object of the same scale as the vehicle or larger, the operator could miss that object due because there are few distinguishing features between large objects and radar malfunctions.

6.2.5 Algorithmic Assistance

During the early stages of this research, an interface concept was proposed that featured many independent artificial intelligence agents assisting the operator. These agents would be designed to hunt for very specific conditions and alert the user when they were confident that specific conditions were met.

For instance, it was observed that a small fraction of the scenario data featured very strong radar returns, far stronger than any noise or clutter. This subset of the scenarios, which was approximately 10% of the total detections, were entirely real objects and no false alarms. An artificial intelligence agent could easily be designed to threshold on the peak strength of a signal and alert the user that this signal was almost guaranteed to be a real object. This strategy is an example of role allocation. Splitting the allocation of responsibility to humans and artificial intelligence agents has been highly successful in the past in other domains [32].

Other types of specific situations for which artificial intelligence could assist the operator include:

- Low navigation confidence -- GPS, IMU, and sensor registration data can diverge, indicating that one or more sensors are not accurate. The likelihood that the system correctly understands its location is assessed with distance metrics comparing the predicted location of the vehicle from IMU data with the best radar registration location.

- Individual sensor element hardware damage -- Radar arrays include many components that can fail in isolation, sometimes only under specific thermal conditions. These conditions are often hard to detect, and lead to specific graphical artifacts which could be identified by automation.

- Short-duration signals -- Short duration signals appearing on one or more radar elements covering only a few samples are a common indication of noise that should be ignored. Noise across all the sensors at once is likely caused by pitching motion.

- Specific object pattern recognition -- In the general case, Gopher cannot rely upon a library of known signals. However customized agents could hunt for common objects relevant to a particular application. For instance, a particular object used in this data set had a very specific three-peak radar cross section. That object would be a useful object to focus on for automated assistance since it proved troublesome for users far more frequently than other object types. This technique presents additional infrastructure challenges and would require more frequent system updates, though the benefits could potentially outweigh the costs.

- Specific macroscopic patterns -- Patterns such as four depressions in the ground spaced exactly the same distance apart as the wheelbase of a truck are likely an indication of a prior vehicle's passage, not four potential objects. In the data used for this research, a vehicle frequently stopped in mud and sunk into the ground. This left a specific four-peaked signal that could be used as the basis for automated classification.

As mentioned in Chapter 3, expert opinions indicated that the z-dimension did not provide much useful added information relevant to the signal classification process. Therefore to simplify the interface, the 3D voxel radar data was condensed into a flat plane by summation in the z-dimension. This decision was justified by the goal of reducing operator overload. There is no reason that artificial intelligence, which does not suffer from cognitive overload, should be denied this information. Future research should address how the z-dimension could be used to help identify buried objects and eliminate surface artifacts.

6.2.6 Committees and Self-Awareness

This research examined collective intelligence observed in subjects operating in isolation. The results demonstrated large gains in performance compared to individuals. Further research is needed to examine how self-awareness of a subject's role in the voting process might influence his or her behavior. If an operator knows that his or her vote

might not match the final decision, it might affect his or her confidence. This might diminish the effectiveness of the collaborative process when individual act in isolation.

6.2.7 Higher Levels of Automation

Chapter 3 identified possibly levels of automation for the components of the Gopher system. It is notable that the human operator does not provide any core capabilities of the system by their physical presence. Particularly for applications in hazardous environments, it therefore makes sense to remove the human from the vehicle entirely and operate the system remotely. Transitioning from a human-in-the-environment system to a remotely-piloted system brings many human factors issues which much be addressed. This type of situation has been studied in depth for remotely-piloted aircraft systems, and that body of research can be applied to a remotely-piloted Gopher system.

6.2.8 Demographics

This study revealed differences in the performance of different demographic groups. Gender was an interesting category that demonstrated significant differences between men and women. Future studies should examine the performance of various groups of people to better answer questions about ideal operator candidate demographics. In addition to demographic data, various testing metrics such as personality types and ability scores may be correlated with subject performance. If evidence of a correlation between high scores in cognitive tasks and operator performance could be found, it would make recruiting ideal operators easier and more successful.

Appendix A

Experiment Materials

The following documents detail the paperwork used with each subject, including the consent process and survey materials.

GPR Experiment Checklist

- ☐ Assign subject numbers _____ and _____ .
- ☐ Determine if subject should be even first or odd first: _______ first.
- ☐ Write subject number on all 4 forms.
- ☐ Obtain consent
- ☐ Store consent form
- ☐ Present pre-experiment questionnaire
- ☐ Store pre-experiment questionnaire
- ☐ Show training slides for first number
- ☐ Ask about questions
- ☐ Start app with first subject number
- ☐ Review performance
- ☐ (Review performance if required after remedial training)
- ☐ Start subject on experiment
- ☐ Present post-experiment questionnaire
- ☐ Store post-experiment questionnaire
- ☐ Show training slides for second number
- ☐ Ask about questions
- ☐ Start app with second subject number
- ☐ Review performance
- ☐ (Review performance if required after remedial training)
- ☐ Start subject on experiment
- ☐ Present post-experiment questionnaire
- ☐ Store post-experiment questionnaire
- ☐ Give compensation
- ☐ Explain competition winner will be emailed about a month later
- ☐ Debrief and answer any final questions
- ☐ Thank subject

CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

Ground Penetrating Radar Development (Program 1958)

You are asked to participate in a research study conducted by Dr. Mary L. Cummings, from the Aeronautics/Astronautics Department at the Massachusetts Institute of Technology (M.I.T.) as a part of Paul W. Quimby's thesis research. You were selected as a possible participant in this study because of your interest in improving displays for ground-penetrating radar data. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.
Your may be withdrawn from the research if your vision is worse than 20/25 when corrected with glasses or contacts or if you do not understand English well enough to follow the instructions of this experiment.

- **PURPOSE OF THE STUDY**

This experiment concerns the interface for a vehicle-mounted Ground-Penetrating Radar (GPR) array. GPR is used in a variety of applications including civil engineering, utility maintenance, demining, archaeology, and the earth sciences. Because GPR signals are often challenging to interpret, we hope to learn how to make more effective use of GPR technology by focusing on how to support a human interacting with an automated GPR detection system. In particular, this study attempts to discover how to design a better human interface by observing how different interfaces affect human performance.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:
- Participate in a training session to familiarize yourself with the display and test conditions. (15 minutes)
- Use the interface to classify example scenarios. (30 minutes)
- Answer survey questions about your interaction with the system. (15 minutes)

All of these steps will occur either in the MIT Rooms 35-220, 37-301 or in the Lincoln Laboratory Room S2-330, depending on your schedule.

- **POTENTIAL RISKS AND DISCOMFORTS**

There are no foreseeable risks, discomforts, or inconveniences in participating in this experiment.

- **POTENTIAL BENEFITS**

You will not personally benefit from participating in this study. Instead you will help society by helping scientists and engineers learn about human interaction with automated systems. The benefits from greater understanding of human-computer interaction include more user-friendly interfaces and more effective, safer, and faster automated systems.

- **PAYMENT FOR PARTICIPATION**

You will be paid $15 for your participation in this study, which will be paid upon completion of your debrief. Should you elect to withdraw during the study, you will be compensated for your time proportionally to one hour.

To encourage you to perform as best as you can if you achieve the best performance of all the experiment participants will be given a $100 Best Buy Gift Card. This prize will be distributed at least one month after your participation today. You will be contacted by the experimenter if you are selected as the top performer.

- **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

Your performance and survey responses in this study will only be recorded by your subject number, which will not be publicly linked to your name so your participation in this research is essentially anonymous. The only places your name and/or subject number will be recorded are this document and as required by MIT for the purposes of documenting consent to this experiment. This document and any copies will be kept in a locked storage location according to MIT policies for retaining such records.

- **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact

Principal Investigator
Mary L. Cummings
77 Massachusetts Ave.
33-311
Cambridge, MA 02139
ph. 617 252 1512

Student Investigator
Paul W. Quimby
77 Massachusetts Av.
35-220
Cambridge, MA 02139
ph. 978 844 4057

- **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- **RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

____

Name of Subject

____

Name of Legal Representative (if applicable)

______________________________________ ______________

Signature of Subject or Legal Representative Date

SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

______________________________________ ______________

Signature of Investigator Date

Pre-experiment Survey

Subject Number ________

Please answer the following questions about yourself. Should you wish to skip a question, you may do so without jeopardizing your participation in this study.

1. Gender: ______

2. Age: ______

3. Occupation: ____

4. Military experience: Yes No (circle one)
   If yes, which branch: ___________
   If yes, years of service: ___________

5. Are you colorblind? (circle one)

   Yes No If yes, which colors: ________________________________

6. How frequently do you use a touchscreen device such as an iPad or iPhone?

   Never A few times ever A few times a year Daily

7. Have you studied or used radar devices before? (circle one)

   Yes No If yes, please explain:

   ____

   ____

   ____

8. How frequently do you play video games currently? (circle one)

Rarely Once a month Weekly A few times a week Most days

9. When you make decisions are you a conservative or risky decision maker? (circle one)

Very conservative Conservative Neutral Risky Very risky

Post-experiment Survey 1

Subject Number _______

Please answer the following questions to help us improve our interfaces. Should you wish to skip a question, you may do so without jeopardizing your participation in this study.

1. What percentage of detections do you think you correctly answered? _______ %
2. Which were the easier examples to classify? Why were they easier?
3. Which were the more challenging examples? Why were they more challenging?
4. Do you have any other comments for me?

Post-experiment Survey 2

Subject Number _______

Please answer the following questions to help us improve our interfaces. Should you wish to skip a question, you may do so without jeopardizing your participation in this study.

1. What percentage of detections do you think you correctly identified using the second interface?

   _______ %

2. For the second interface, which were the easier examples to classify? Why were they easier?

3. For the second interface, which were the more challenging examples? Why were they more challenging?

4. Were there any aspects of the first interface that made your job easier or harder compared to the second interface?

5. Were there any aspects of the second interface that made your job easier or harder compared to the first interface?

6. Which interface do you prefer? (circle one)

   Strongly prefer 1st   Prefer 1st   No preference   Prefer 2nd   Strongly prefer 2nd

7. Do you have any other comments for me?

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Training Slides

The training slides shown to subjects are included in this appendix. Each subject saw both sets of training slides, though the order was counterbalanced.

GPR Training

Odd Subject Numbers
Interface A

1

Overview

- This experiment is about detecting buried objects using ground-penetrating radar mounted on a vehicle.
- Our goal is to design a better user interface that makes it easier to find buried objects.
- Finding objects is difficult because we do not know exactly what to look for and many things may cause noise in the data.

2

- You can help us design a better interface by learning how to use this interface and then testing it for us.
- Over the next several slides, we will explain the display to you and show you examples.
- After we show you the examples, we will give you a chance to practice using the system.

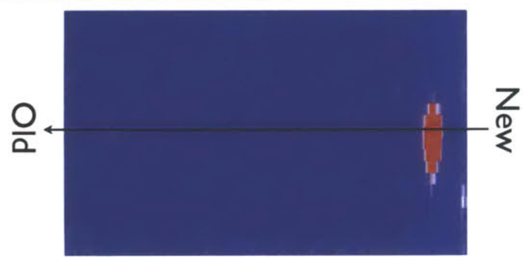
3

- After practicing, we will show you many examples and ask for your opinion.
- Each example will either be real radar data that contains a buried object, or real data that was noise.
- The person who gets the most correct answers will receive a $100 gift card.

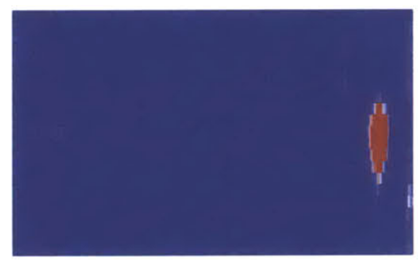
4

- This is NOT an exercise in speed. You can take as long as you like. Your goal is to get the right answer.
- After every ten examples, you will have the chance to take a break before continuing.
- If you do not understand any of the instructions or if you have any questions, please feel free to ask Paul.

- This is an example display.
- Blue values mean "zero". Red values are strong signals. Gray values are in between.



- This display is organized by time.
- The right edge is the newest data.
- The left edge is the oldest data.
- When you are shown the picture, it means the vehicle has stopped moving and this is the state of the display.

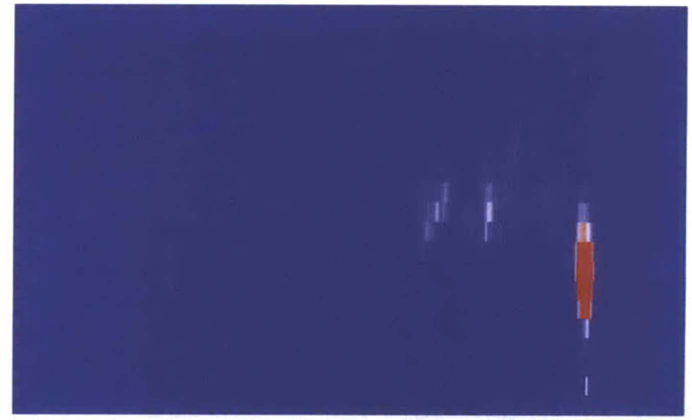


- When the vehicle has stopped, it stops because the automation thinks it found a potential object.
- On the right side of this display you can see a red blob where the system found something and decided to stop.

- The rest of this slideshow is filled with examples of objects and noise so you can learn to tell the difference.
- The next several slides are all OBJECTS.
- The top of each slide tells you which category the example is (either OBJECT or FALSE ALARM).

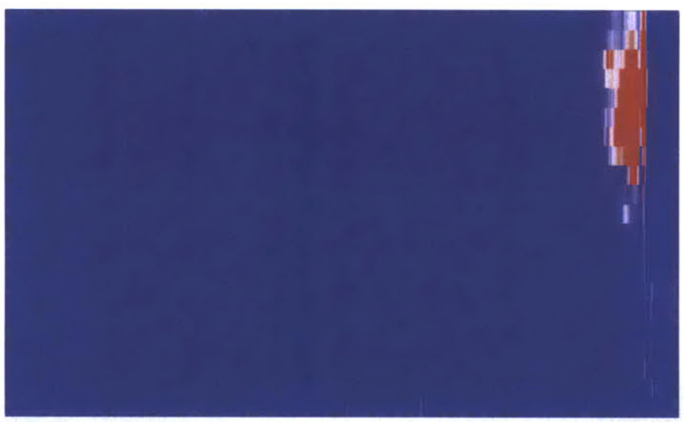
9

OBJECT



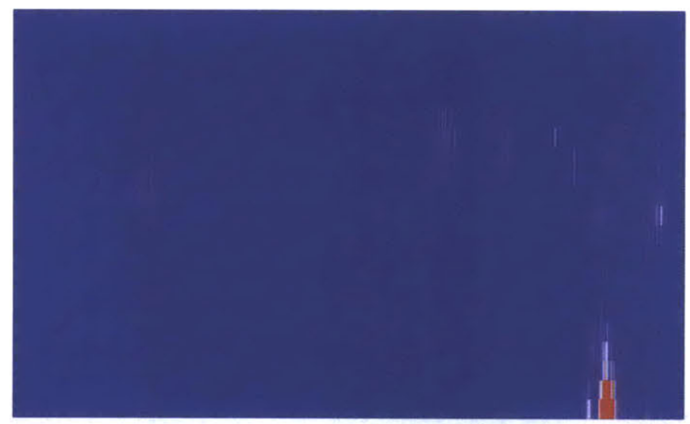
10

OBJECT



11

OBJECT

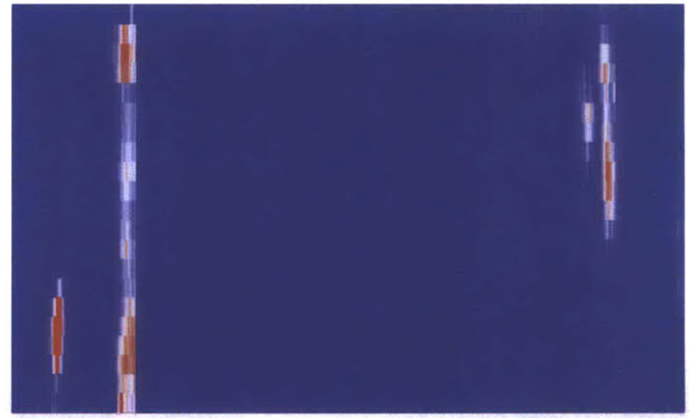


12

OBJECT

13

OBJECT

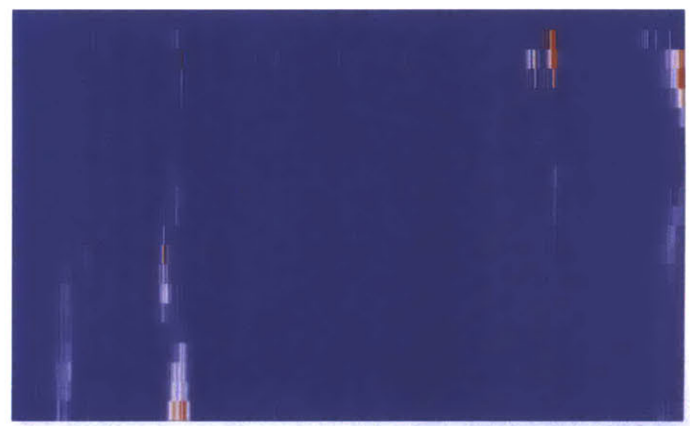


14

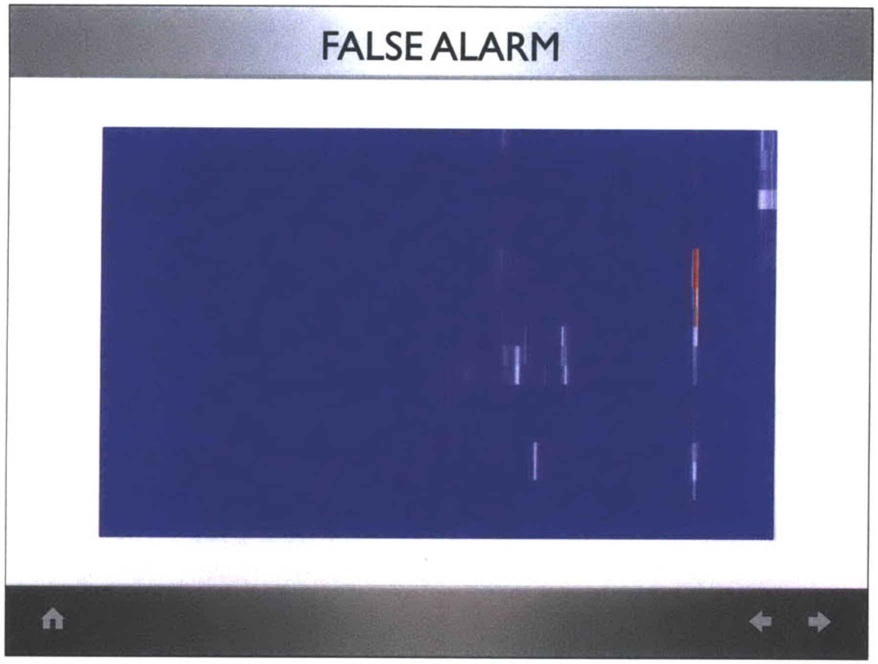
- The next several slides are all FALSE ALARMS.

15

FALSE ALARM

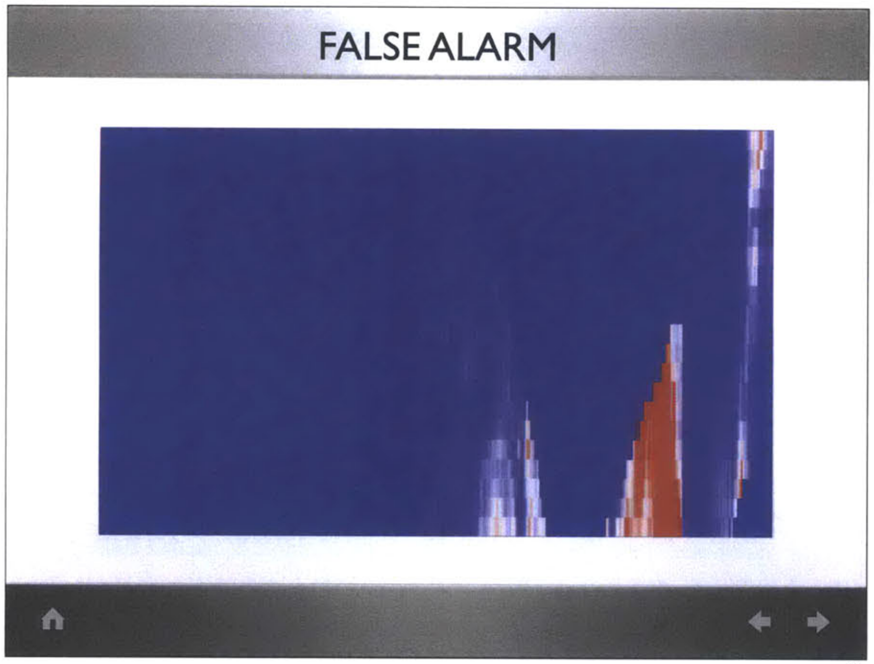


16

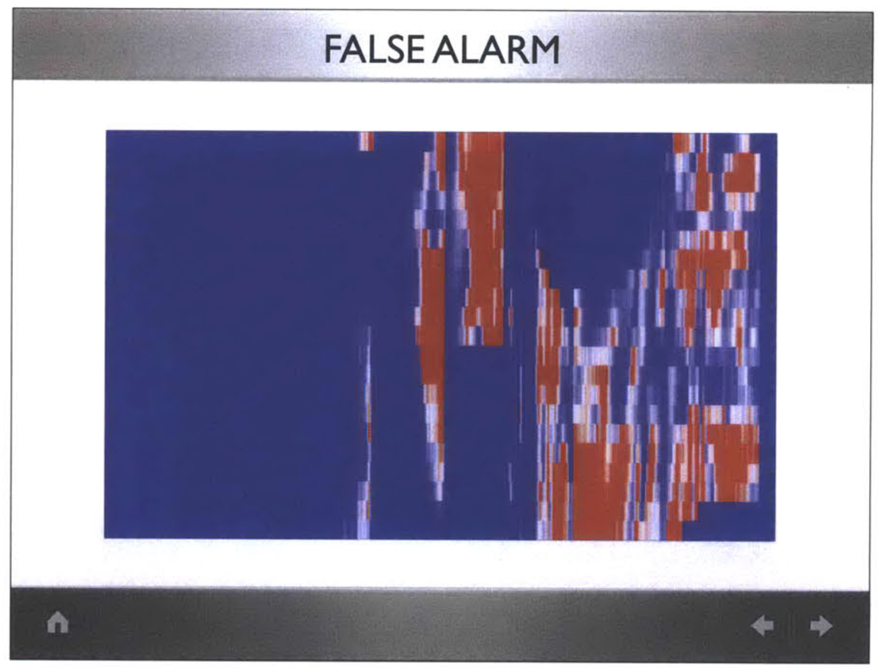


17

18



19



20

- You will now practice classifying detections.
- After each practice example you will be asked to indicate your confidence that you answered correctly.
- After each practice example you will be shown the correct answer.
- Please let Paul know you finished.

GPR Training

Even Subject Numbers
Interface B

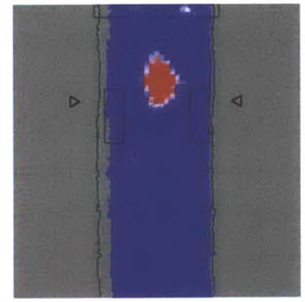
Overview

- This experiment is about detecting buried objects using ground-penetrating radar mounted on a vehicle.
- Our goal is to design a better user interface that makes it easier to find buried objects.
- Finding objects is difficult because we do not know exactly what to look for and many things may cause noise in the data.

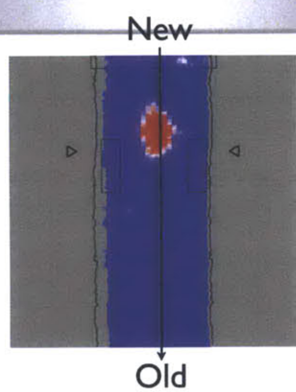
- You can help us design a better interface by learning how to use this interface and then testing it for us.
- Over the next several slides, we will explain the display to you and show you examples.
- After we show you the examples, we will give you a chance to practice using the system.

- After practicing, we will show you many examples and ask for your opinion.
- Each example will either be real radar data that contains a buried object, or real data that was noise.
- The person who gets the most correct answers will receive a $100 gift card.

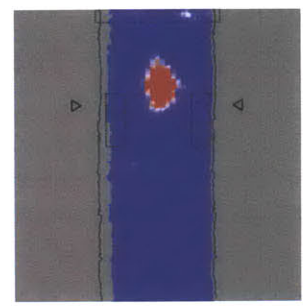
- This is NOT an exercise in speed. You can take as long as you like. Your goal is to get the right answer.
- After every ten examples, you will have the chance to take a break before continuing.
- If you do not understand any of the instructions or if you have any questions, please feel free to ask Paul.



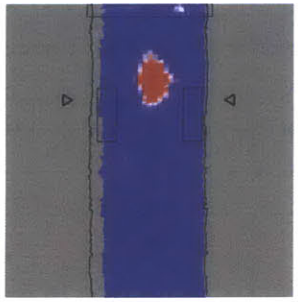
- This is an example display.
- Blue values mean "zero". Red values are strong signals. Gray values are in between.



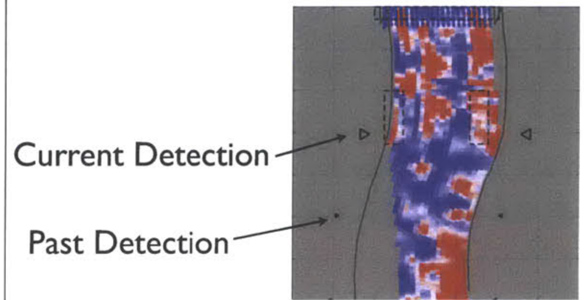
- The vehicle moves forward, "up", so the top edge is the newest data, the bottom edge is the oldest.
- When you are shown the picture, it means the vehicle has stopped moving and this is the state of the display.



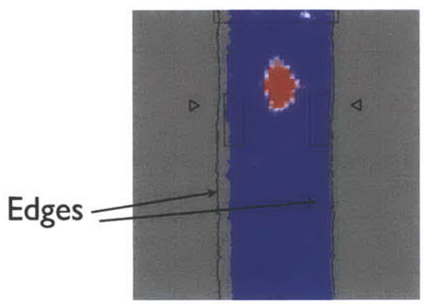
- This display is a top-down map centered on the vehicle.
- The vehicle is always facing "up" on the picture.
- When you are shown the picture, it means the vehicle has stopped moving and this is the state of the display.



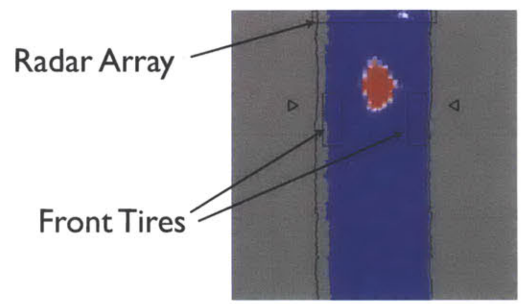
- At the top of this display you can see a red blob where the system found something and decided to stop.
- When you are shown the picture, it means the vehicle has stopped moving and this is the state of the display.



- Big triangles indicate what data triggered the current detection (what made the system stop). You should focus on this signal as opposed to ones further down the display.
- Small triangles indicate previous detections. In all the cases we will show you, these are False Alarms.



- The dark gray area outside the black lines is area without radar data.
- The black lines are the edges of the vehicle's path.

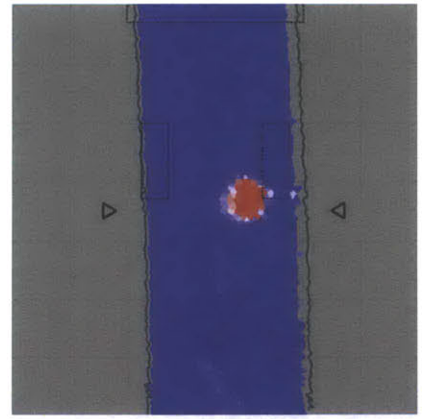


- The box at the top of the display is the radar array.
- The boxes just above the middle of the display are the front tires of the vehicle.
- The rear tires are just below the bottom of the display.

- The rest of this slideshow is filled with examples of objects and noise so you can learn to tell the difference.
- The next several slides are all OBJECTS.
- The top of each slide tells you which category the example is (either OBJECT or FALSE ALARM).

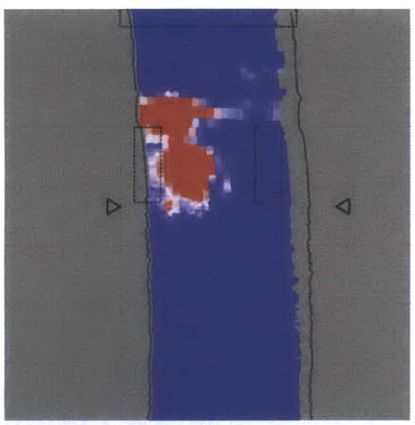
13

OBJECT



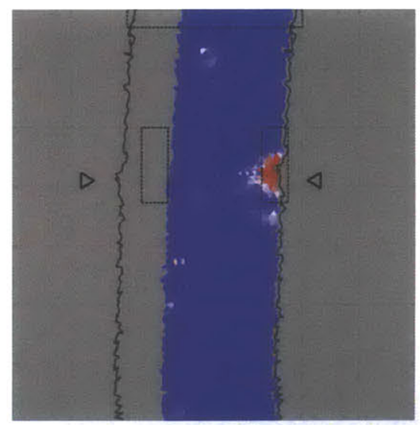
14

OBJECT

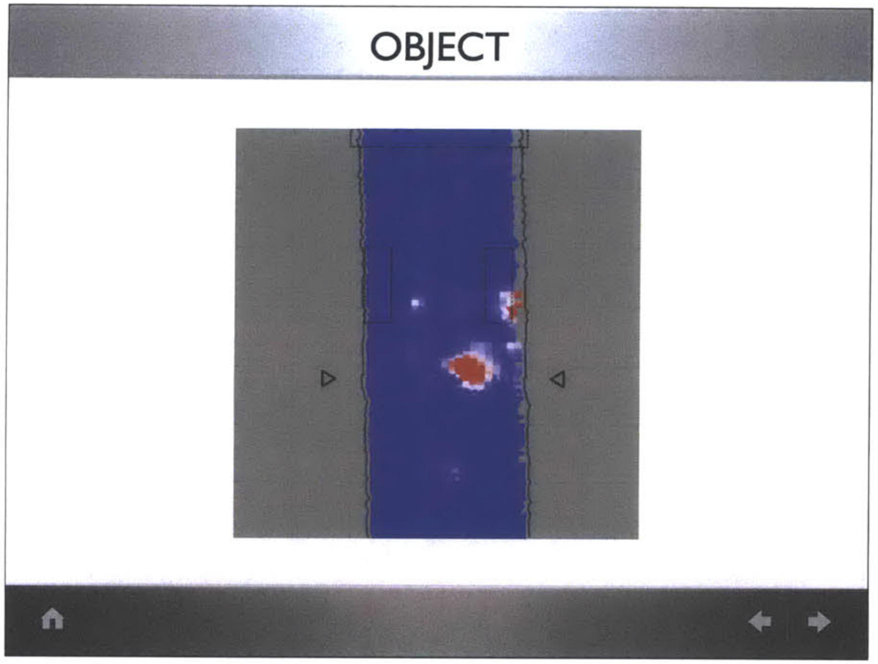


15

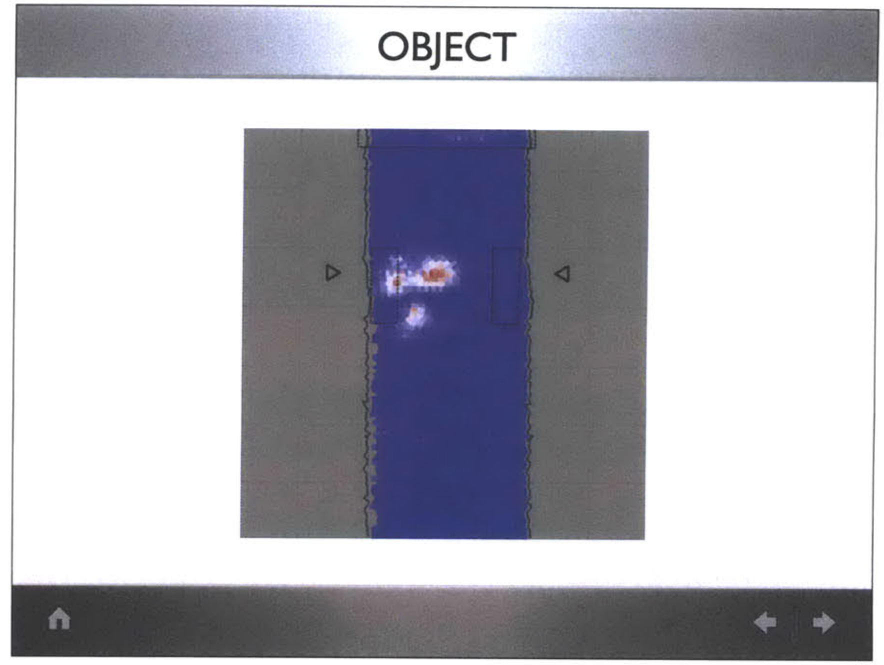
OBJECT



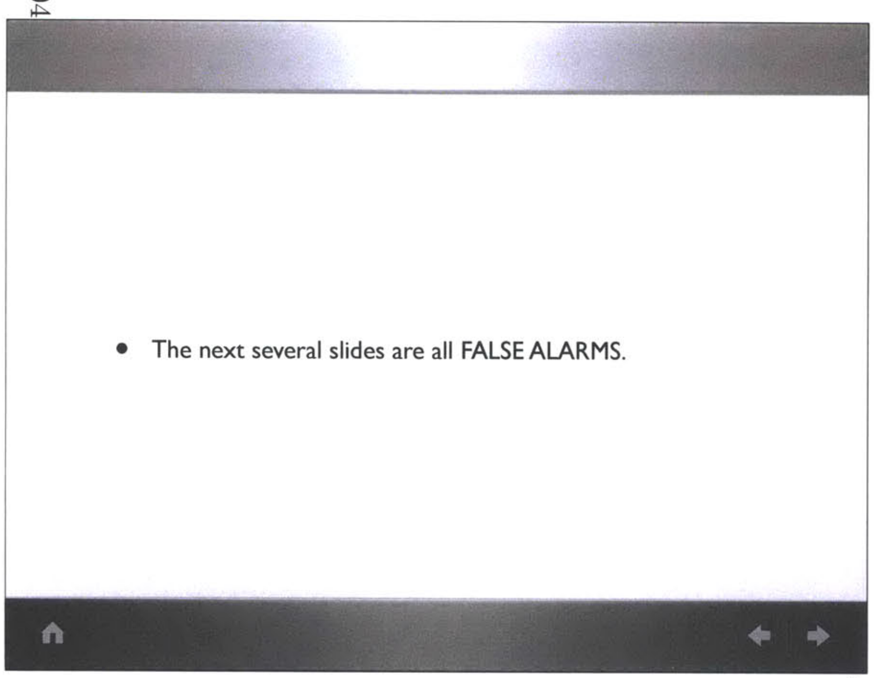
16


OBJECT

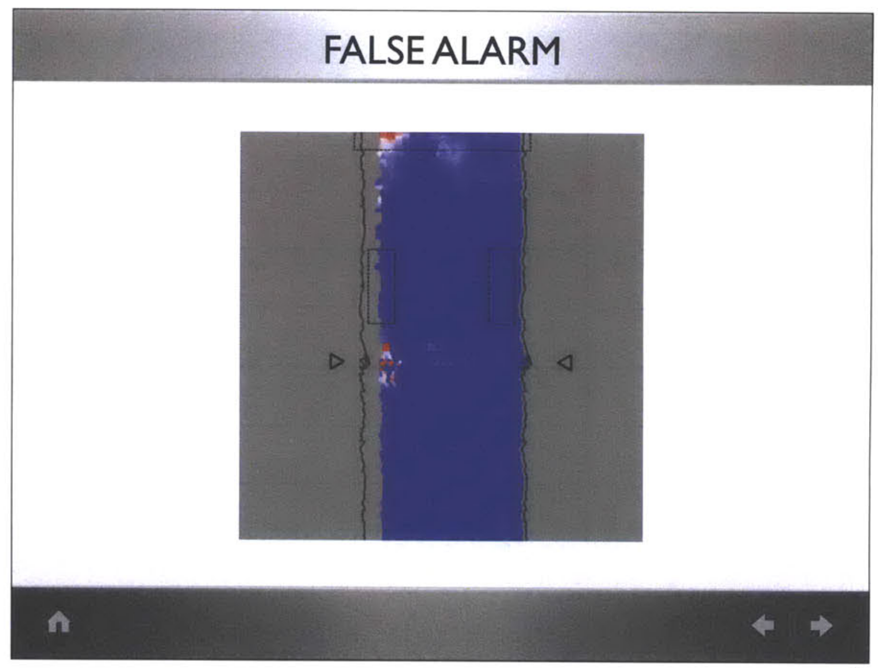
17


OBJECT

18

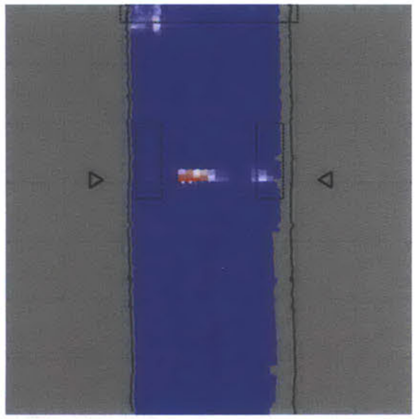

• The next several slides are all FALSE ALARMS.

19


FALSE ALARM

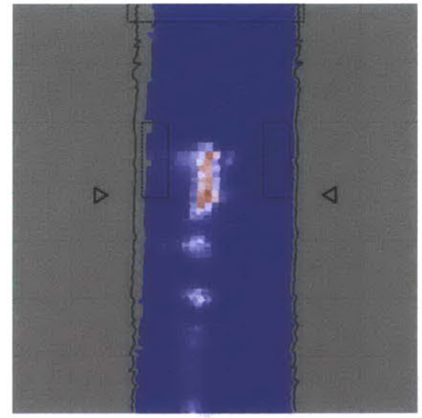
20

FALSE ALARM



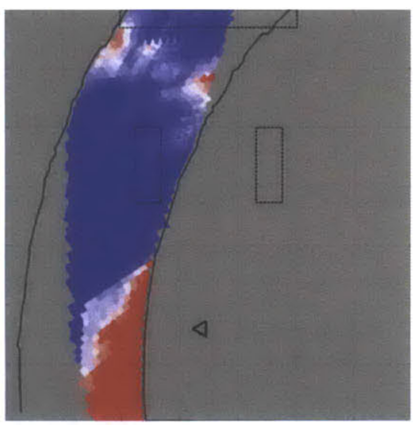
21

FALSE ALARM



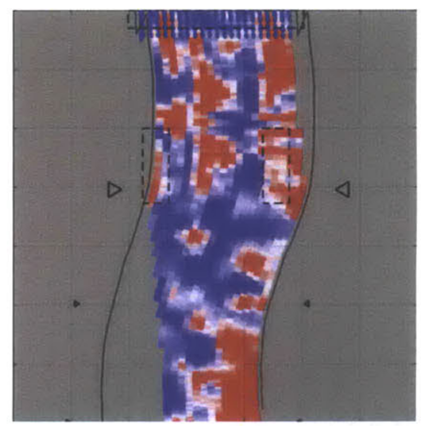
22

FALSE ALARM



23

FALSE ALARM



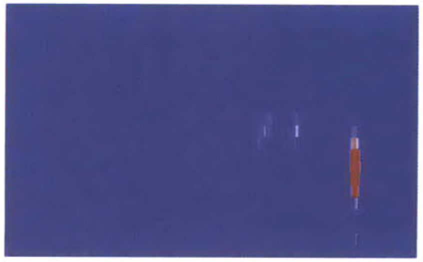
24

- You will now practice classifying detections.
- After each practice example you will be asked to indicate your confidence that you answered correctly.
- After each practice example you will be shown the correct answer.
- Please let Paul know you finished.

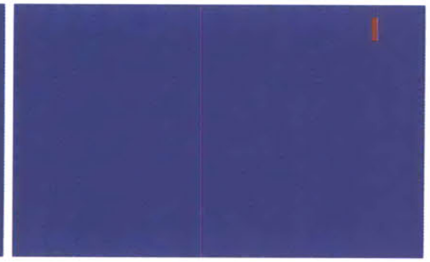
Appendix C

Scenario Library

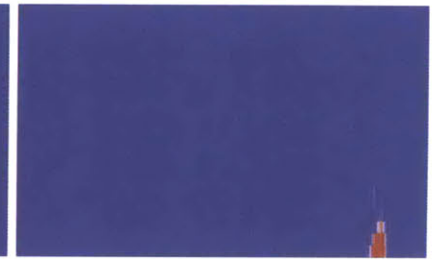
Each scenario used in this experiment is included in this appendix. The scenarios are identified by two numbers in XX-YY format. The correct categorization of each image is included in its label.



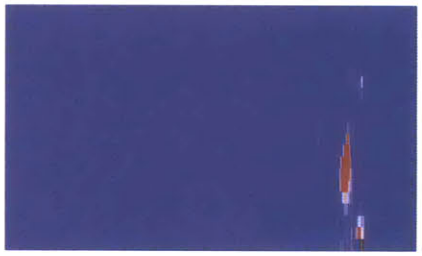
Scenario 01-01, Old, Object



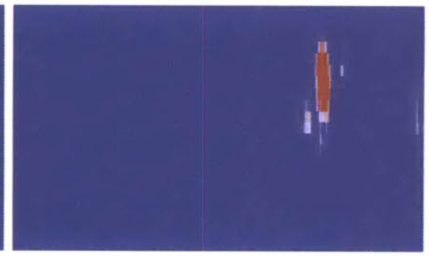
Scenario 01-02, Old, Noise



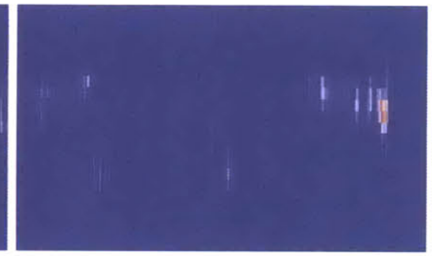
Scenario 01-07, Old, Object



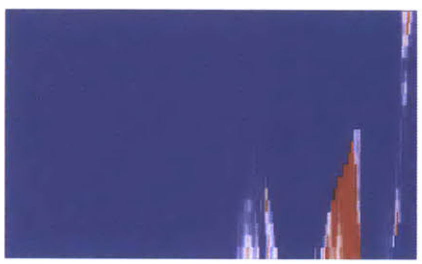
Scenario 01-09, Old, Object



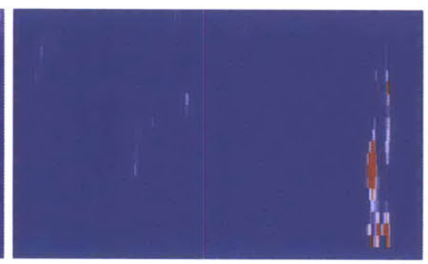
Scenario 01-11, Old, Object



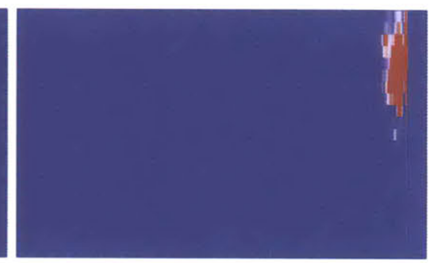
Scenario 01-12, Old, Noise



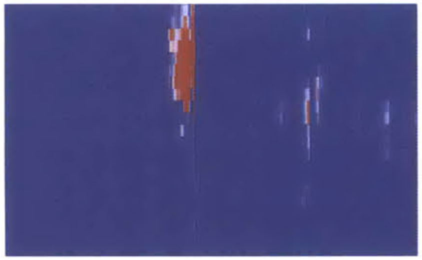
Scenario 01-13, Old, Noise



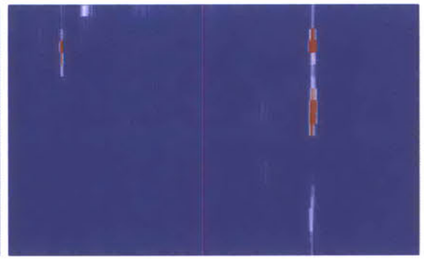
Scenario 02-07, Old, Object



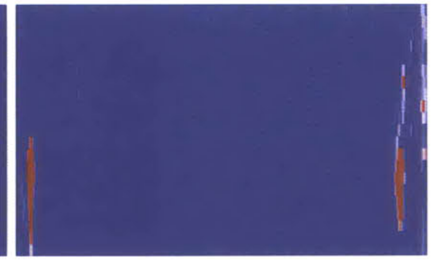
Scenario 02-08, Old, Object



Scenario 02-09, Old, Noise

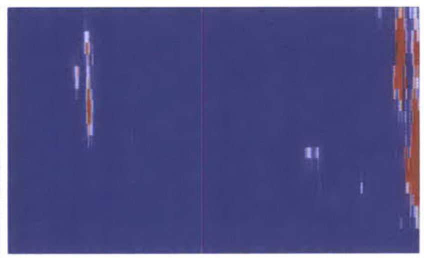


Scenario 02-12, Old, Noise

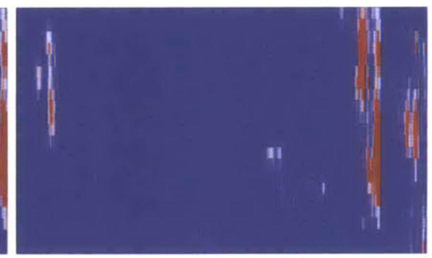


Scenario 03-03, Old, Object

Scenario 03-14, Old, Object

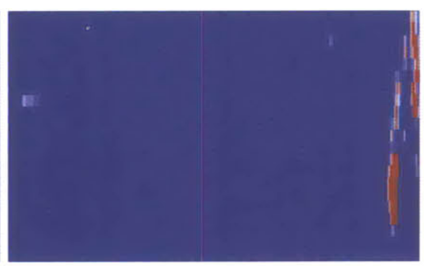


Scenario 03-15, Old, Object

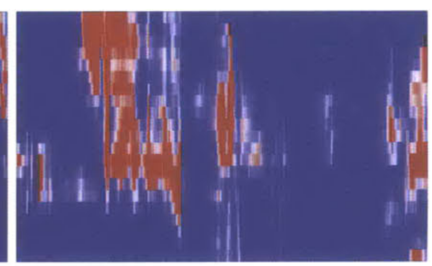


Scenario 03-17, Old, Noise

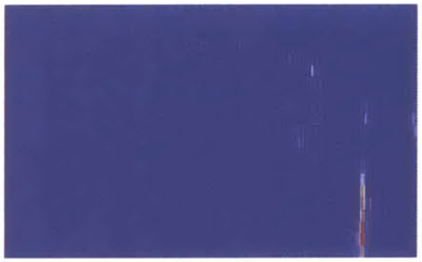
Scenario 03-31, Old, Noise



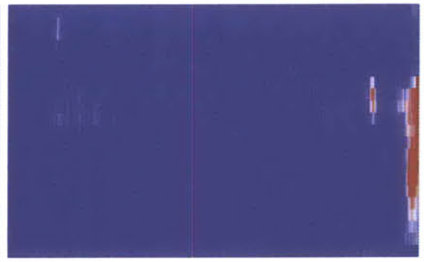
Scenario 04-03, Old, Object



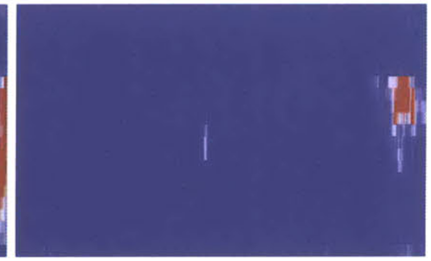
Scenario 04-22, Old, Object



Scenario 04-26, Old, Object

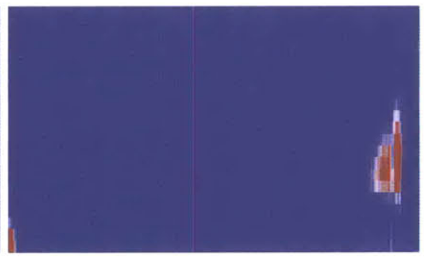


Scenario 04-32, Old, Object



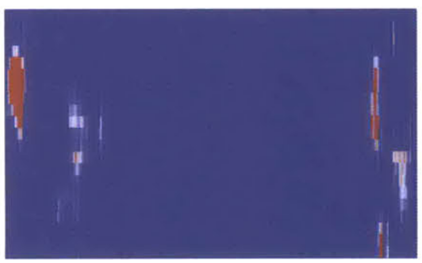
Scenario 05-21, Old, Noise

Scenario 07-02, Old, Object

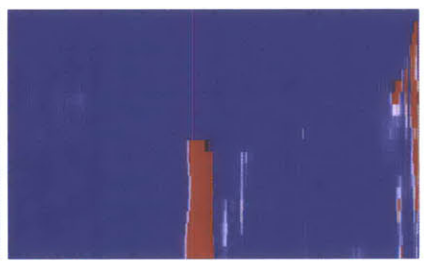


Scenario 07-13, Old, Object

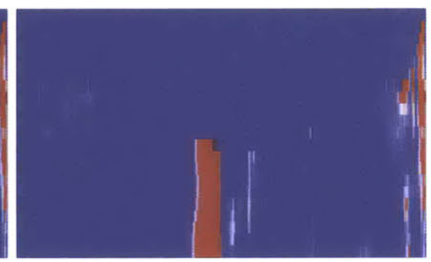
Scenario 07-16, Old, Object



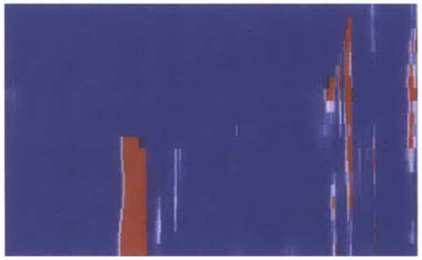
Scenario 07-21, Old, Noise



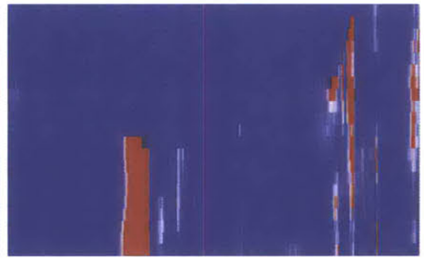
Scenario 07-46, Old, Noise



Scenario 07-47, Old, Noise



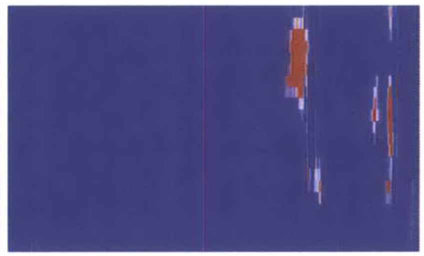
Scenario 07-48, Old, Noise



Scenario 07-49, Old, Noise

Scenario 08-13, Old, Object

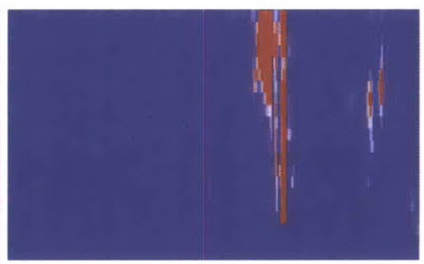
Scenario 08-16, Old, Object



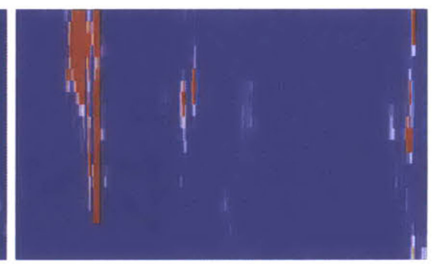
Scenario 08-17, Old, Noise

Scenario 08-37, Old, Noise

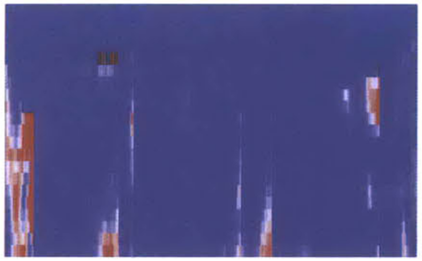
Scenario 09-15, Old, Object

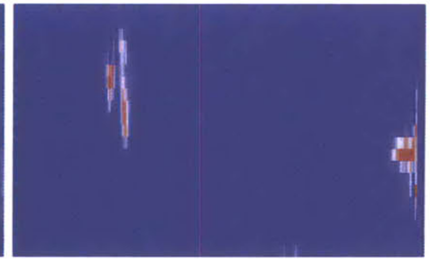


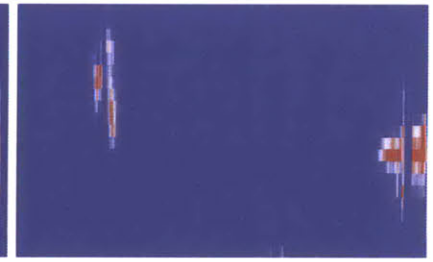
Scenario 09-17, Old, Noise

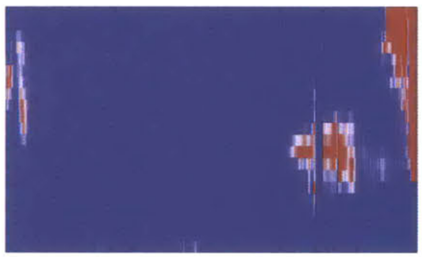


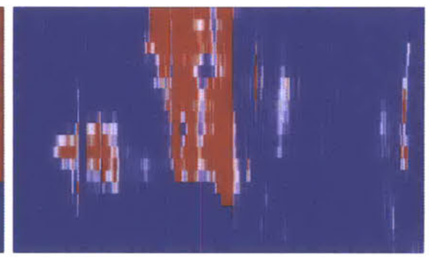
Scenario 09-18, Old, Object

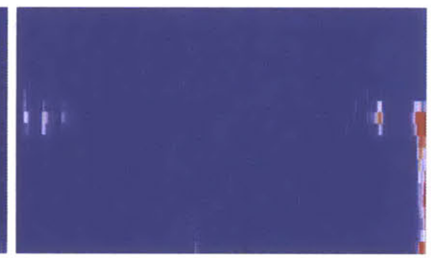

Scenario 09-39, Old, Noise

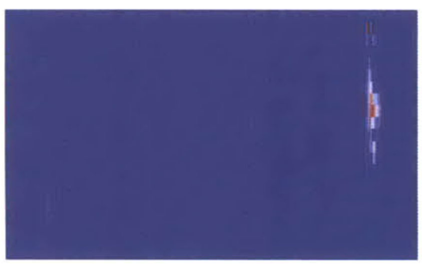

Scenario 10-15, Old, Noise

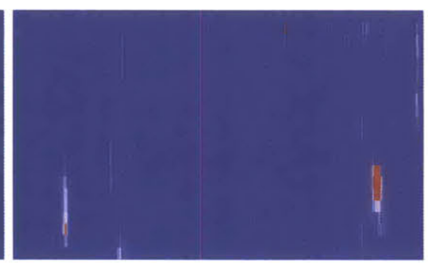

Scenario 10-16, Old, Noise

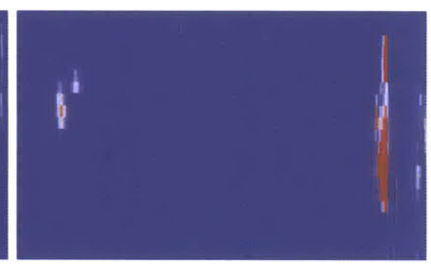

Scenario 10-17, Old, Object


Scenario 10-21, Old, Object

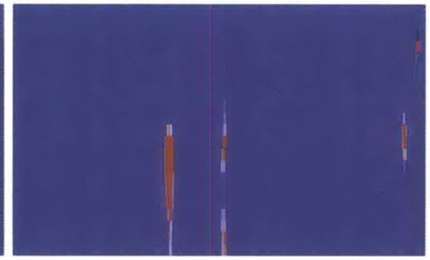

Scenario 10-30, Old, Noise


Scenario 10-34, Old, Object

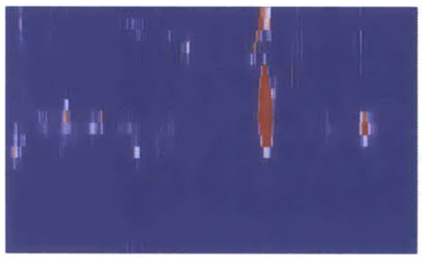

Scenario 10-37, Old, Object

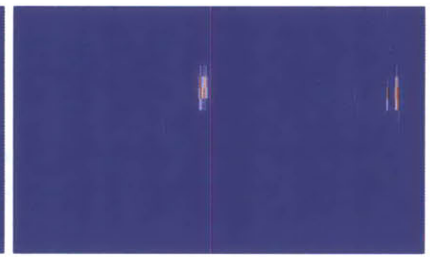

Scenario 10-38, Old, Object

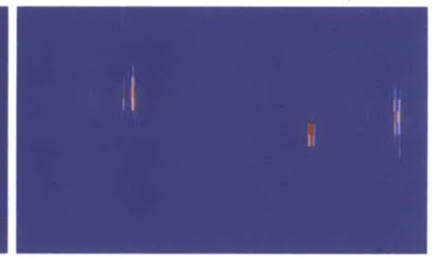

Scenario 11-02, Old, Noise

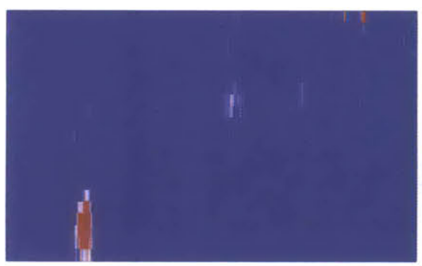

Scenario 11-05, Old, Noise

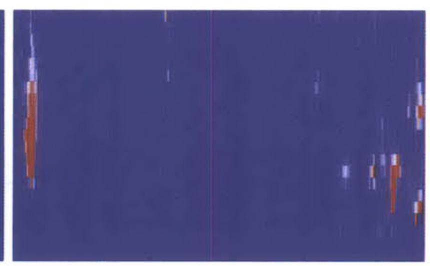

Scenario 11-36, Old, Noise


Scenario 11-38, Old, Noise


Scenario 12-15, Old, Noise

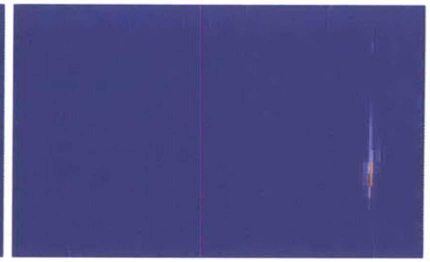

Scenario 12-17, Old, Noise


Scenario 12-25, Old, Noise

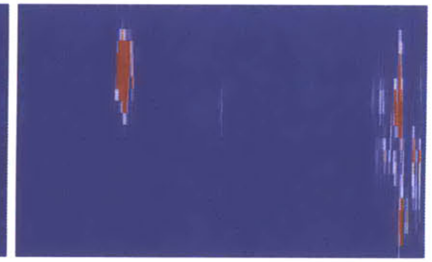

Scenario 12-35, Old, Object


Scenario 12-40, Old, Noise

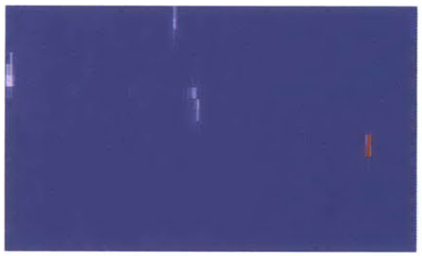
Scenario 13-18, Old, Noise



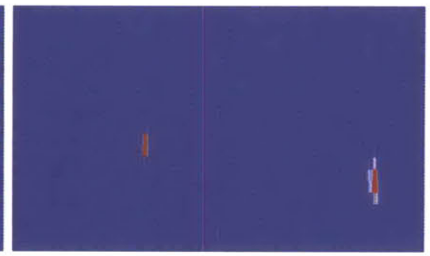
Scenario 13-22, Old, Object



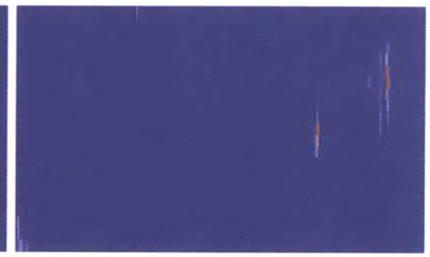
Scenario 14-15, Old, Noise



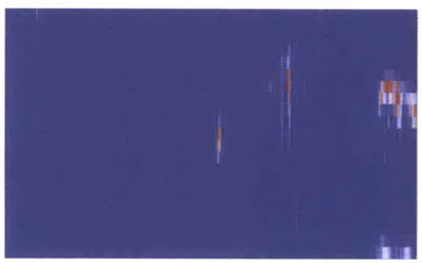
Scenario 14-16, Old, Noise



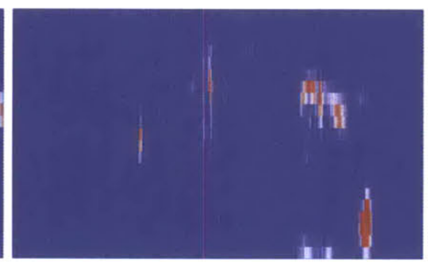
Scenario 14-17, Old, Object



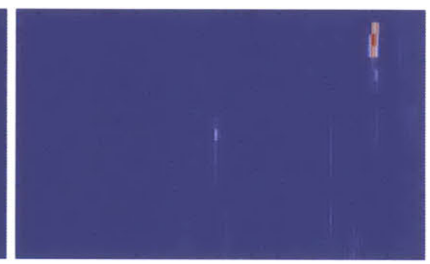
Scenario 14-24, Old, Noise



Scenario 14-25, Old, Noise

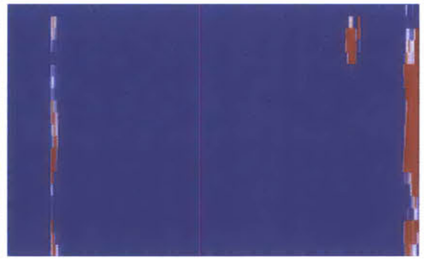


Scenario 14-27, Old, Object



Scenario 14-36, Old, Noise

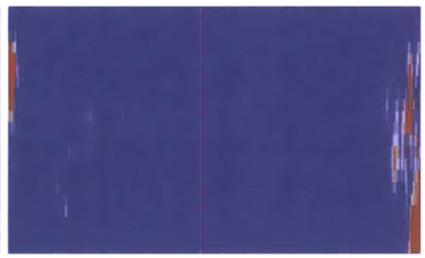
Scenario 15-07, Old, Object



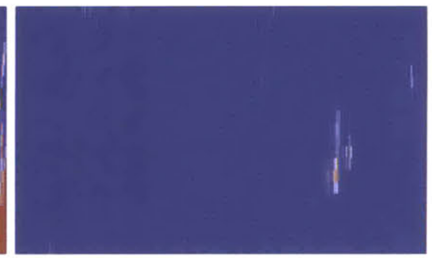
Scenario 16-04, Old, Noise

Scenario 16-05, Old, Noise

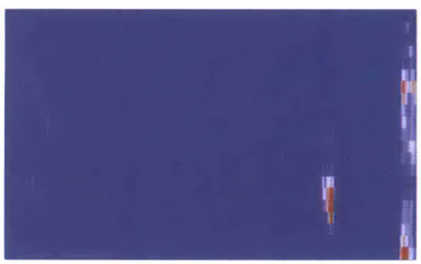
Scenario 16-15, Old, Object



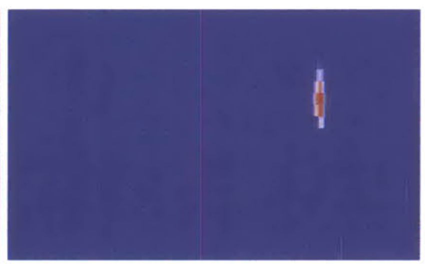
Scenario 16-20, Old, Object



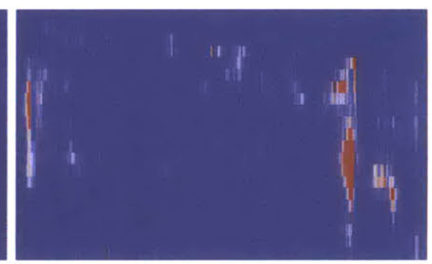
Scenario 16-28, Old, Object



Scenario 16-30, Old, Object

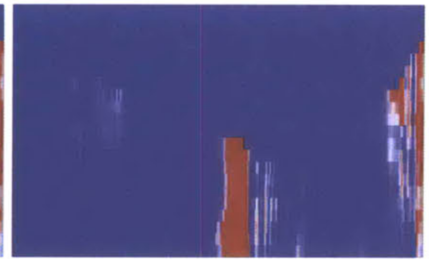


Scenario 16-37, Old, Object

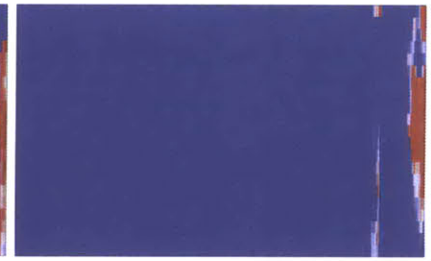


Scenario 16-42, Old, Object

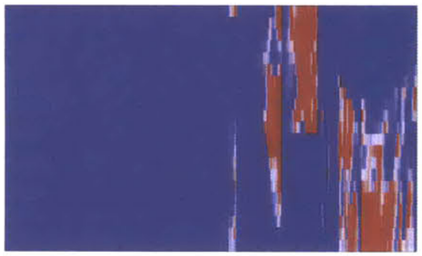
Scenario 16-43, Old, Noise



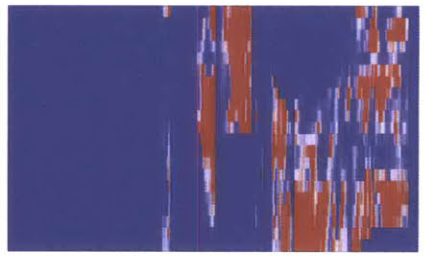
Scenario 16-44, Old, Noise



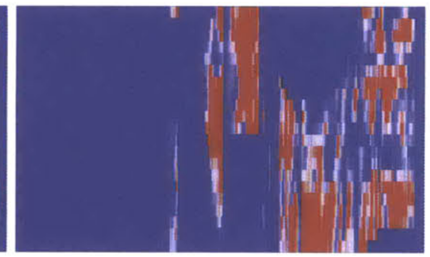
Scenario 17-08, Old, Noise



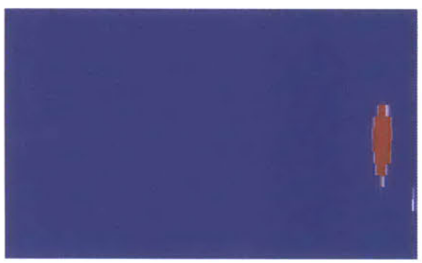
Scenario 17-12, Old, Noise



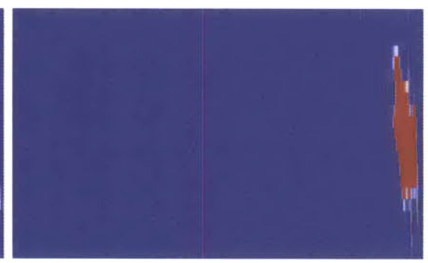
Scenario 17-13, Old, Noise



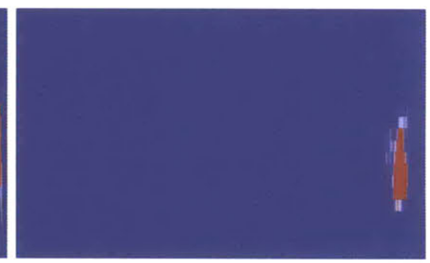
Scenario 17-14, Old, Noise



Scenario 18-05, Old, Object

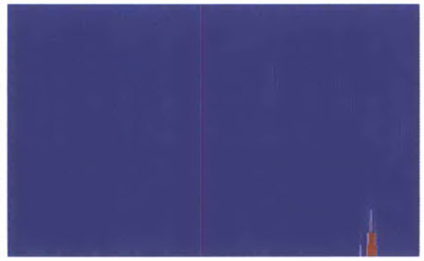


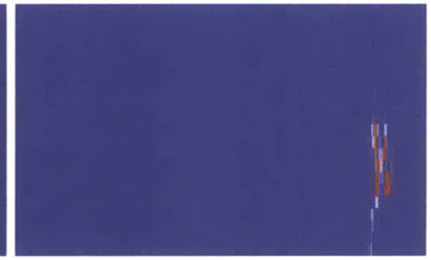
Scenario 18-07, Old, Object



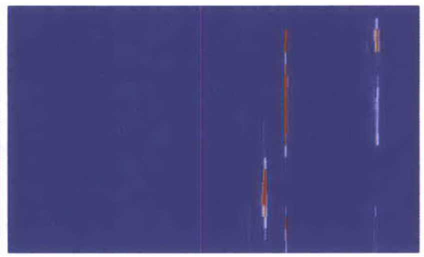
Scenario 18-09, Old, Object


Scenario 19-07, Old, Object

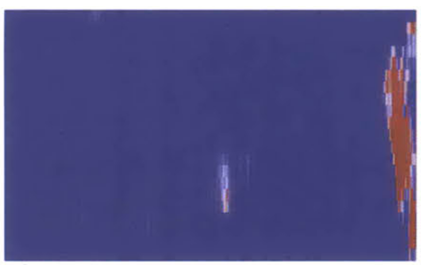

Scenario 21-02, Old, Object

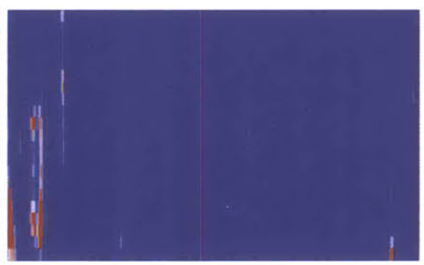

Scenario 21-03, Old, Object


Scenario 21-06, Old, Noise


Scenario 21-09, Old, Noise

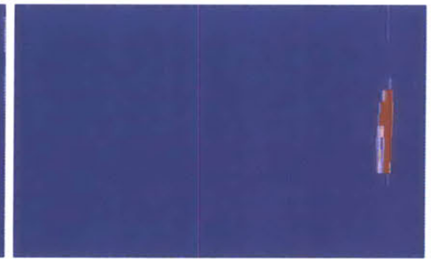

Scenario 21-10, Old, Noise


Scenario 21-11, Old, Object

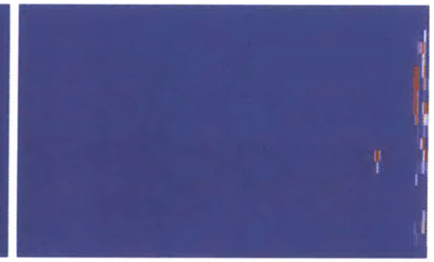

Scenario 21-14, Old, Noise


Scenario 21-16, Old, Object

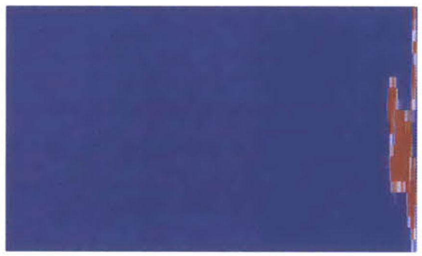
Scenario 21-18, Old, Object



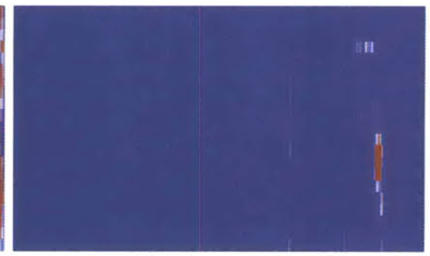
Scenario 22-03, Old, Object



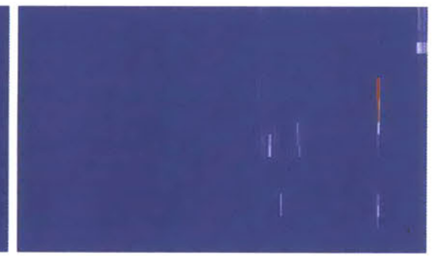
Scenario 22-04, Old, Object



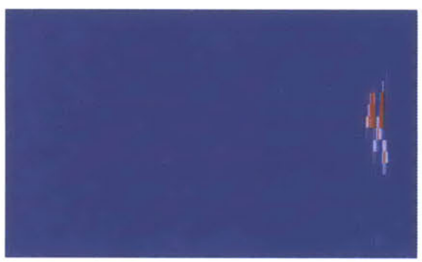
Scenario 22-10, Old, Object



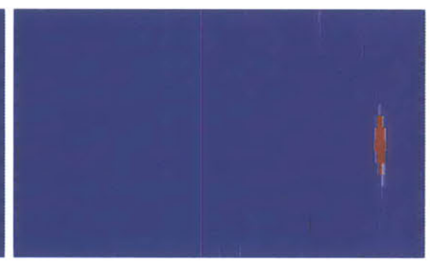
Scenario 22-14, Old, Object



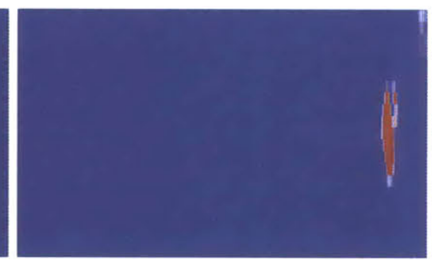
Scenario 22-16, Old, Noise



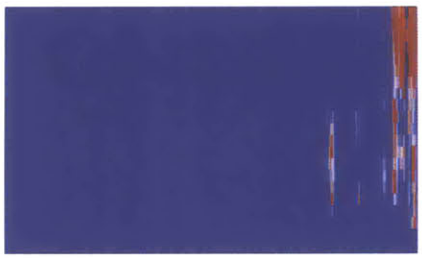
Scenario 22-20, Old, Object

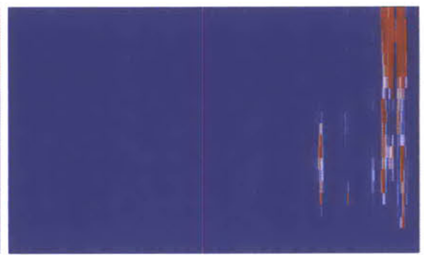


Scenario 22-21, Old, Object

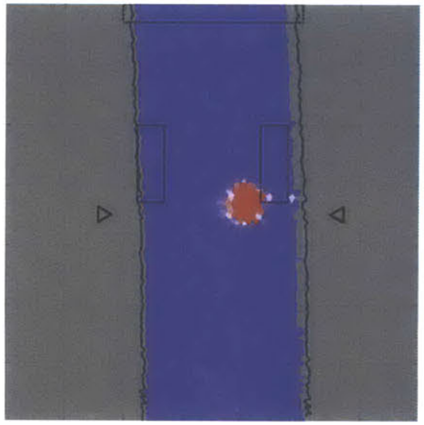


Scenario 22-22, Old, Object

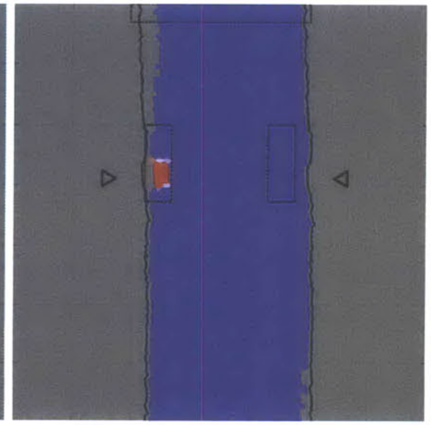

Scenario 25-07, Old, Noise


Scenario 25-08, Old, Noise

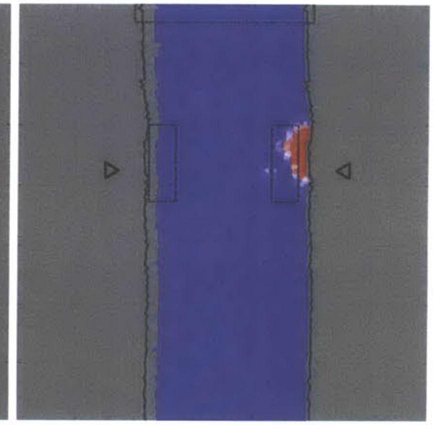

Scenario 25-09, Old, Object



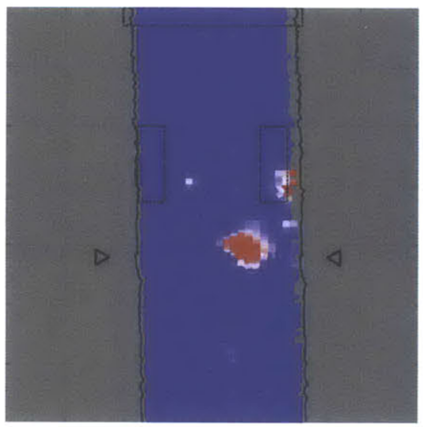
Scenario 01-01, New, Object



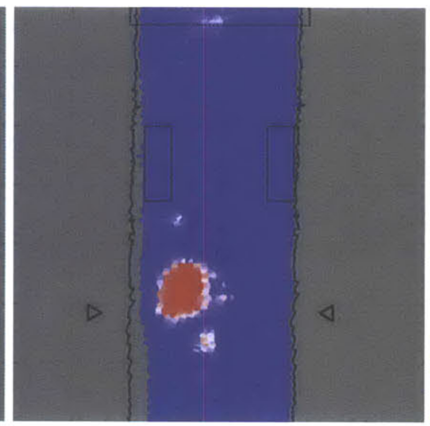
Scenario 01-02, New, Noise



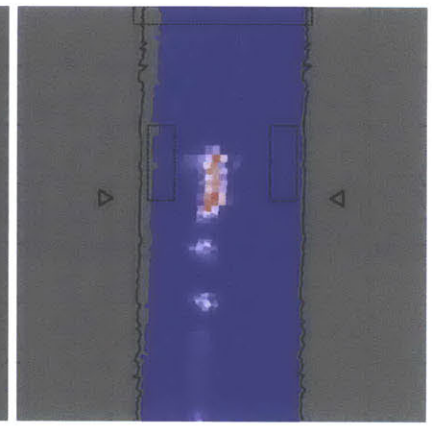
Scenario 01-07, New, Object



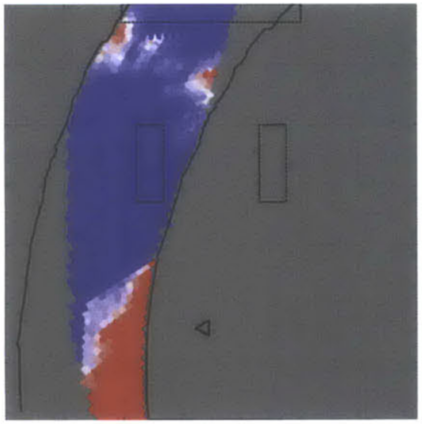
Scenario 01-09, New, Object



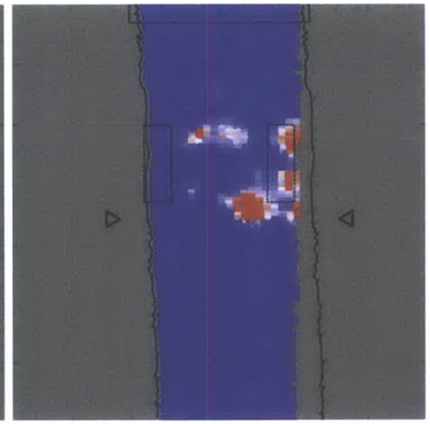
Scenario 01-11, New, Object



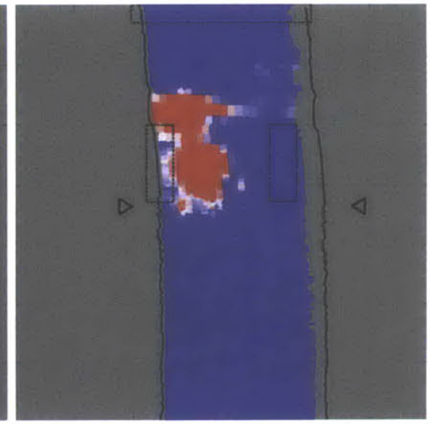
Scenario 01-12, New, Noise



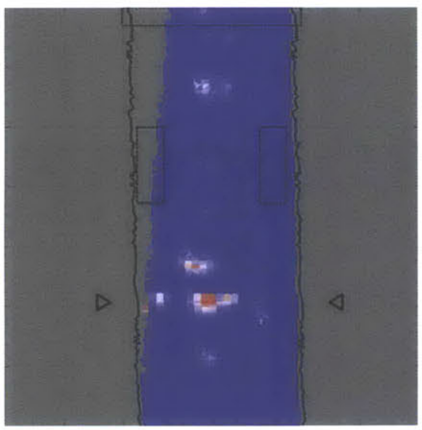
Scenario 01-13, New, Noise



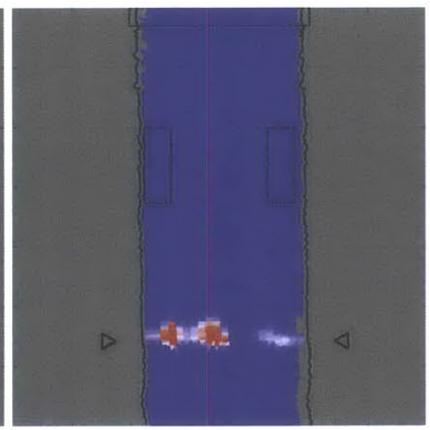
Scenario 02-07, New, Object



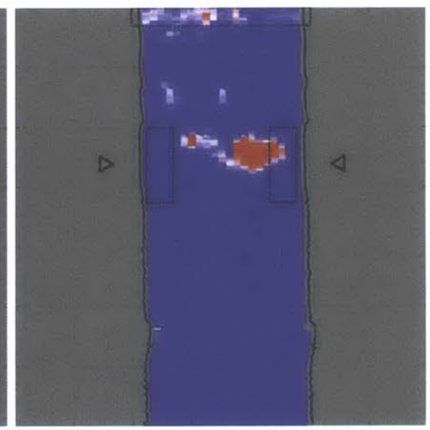
Scenario 02-08, New, Object



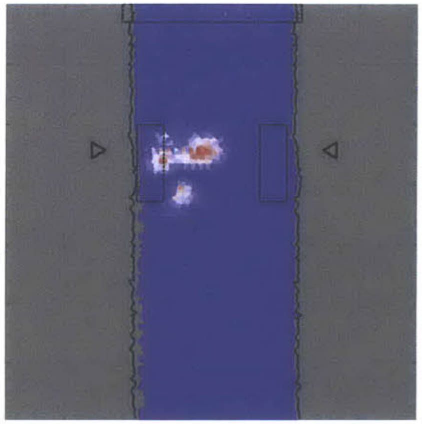
Scenario 02-09, New, Noise



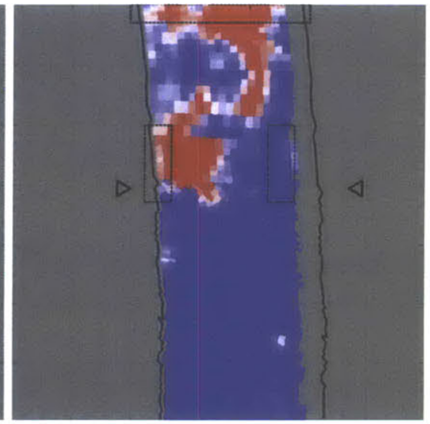
Scenario 02-12, New, Noise



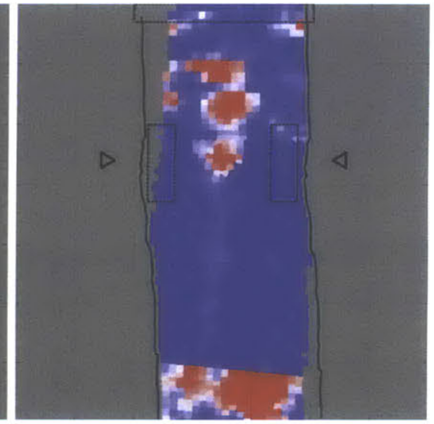
Scenario 03-03, New, Object



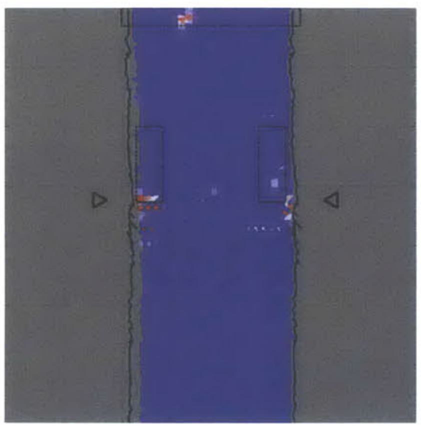
Scenario 03-14, New, Object



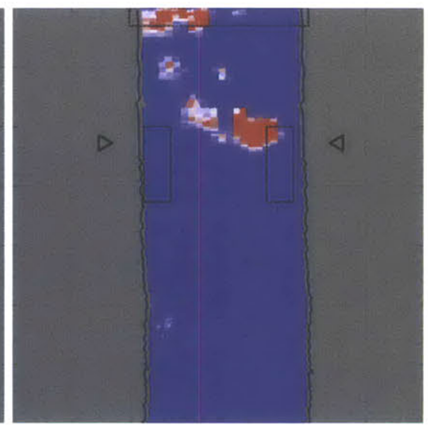
Scenario 03-15, New, Object



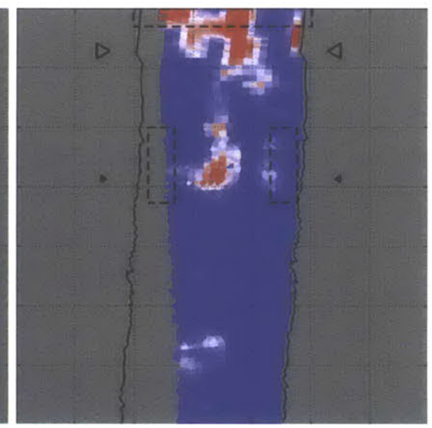
Scenario 03-17, New, Noise



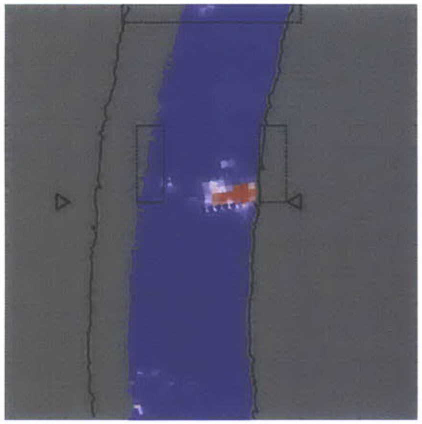
Scenario 03-31, New, Noise



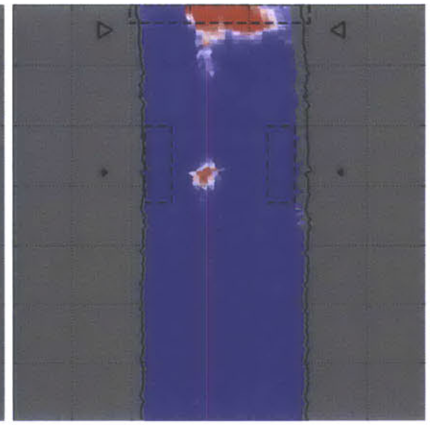
Scenario 04-03, New, Object



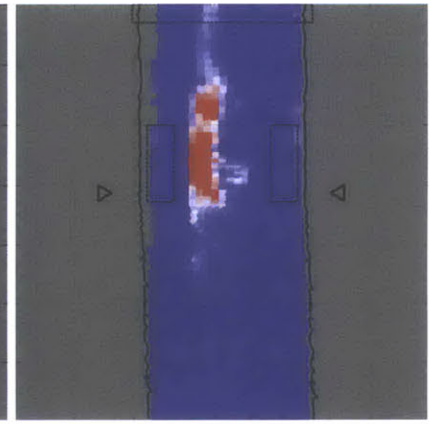
Scenario 04-22, New, Object



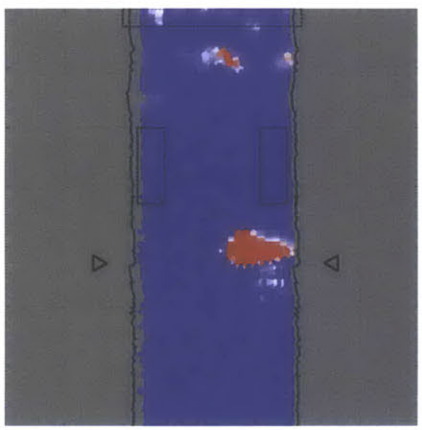
Scenario 04-26, New, Object



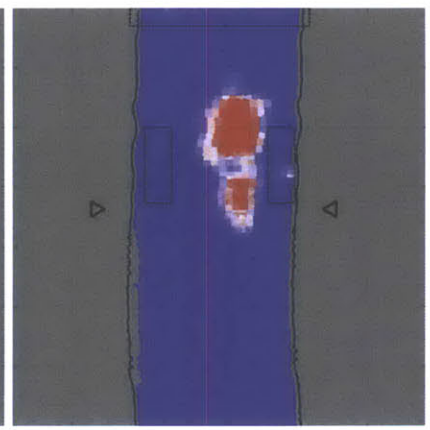
Scenario 04-32, New, Object



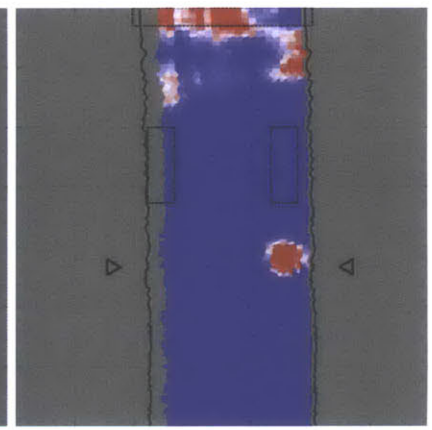
Scenario 05-21, New, Noise



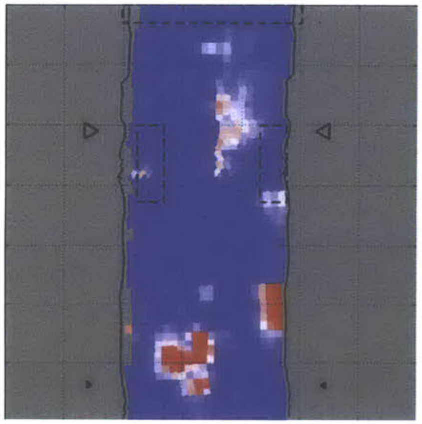
Scenario 07-02, New, Object



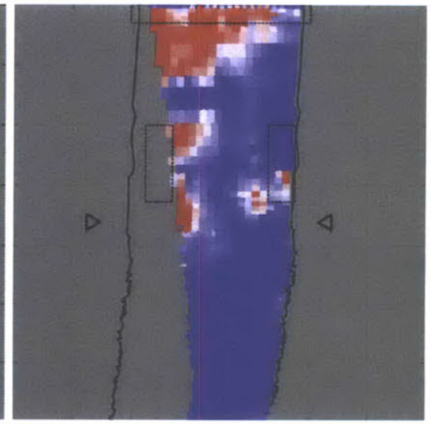
Scenario 07-13, New, Object



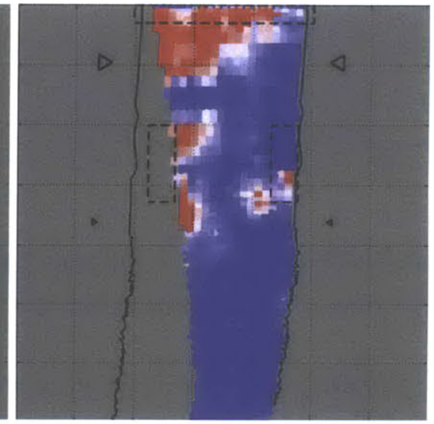
Scenario 07-16, New, Object



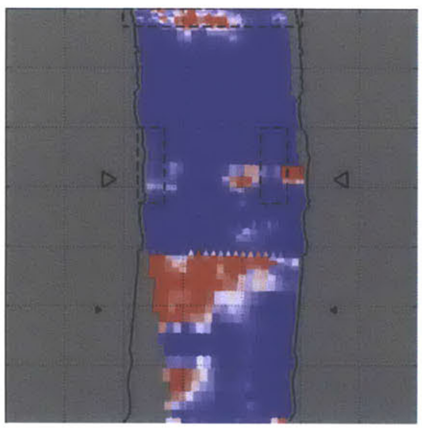
Scenario 07-21, New, Noise



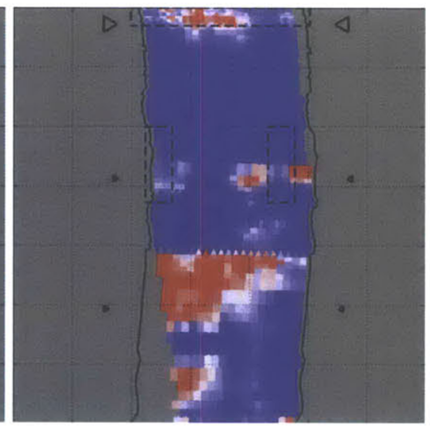
Scenario 07-46, New, Noise



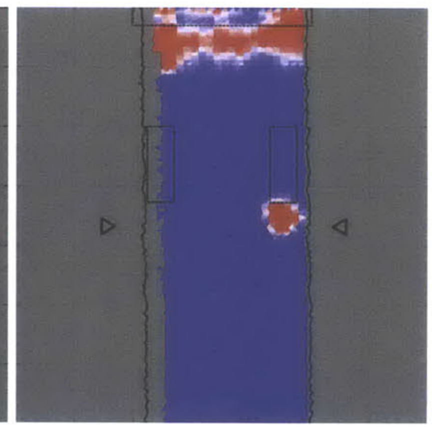
Scenario 07-47, New, Noise



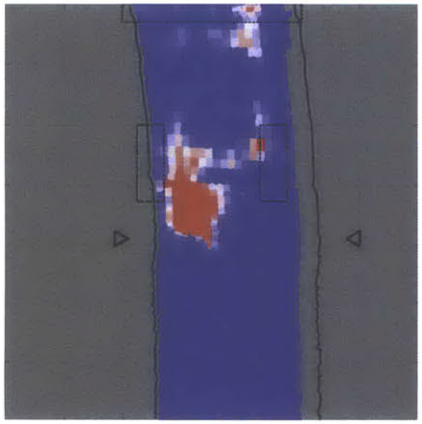
Scenario 07-48, New, Noise



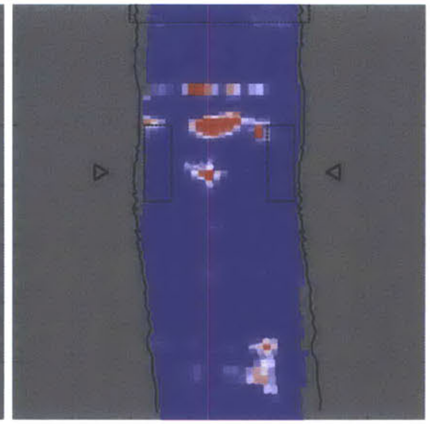
Scenario 07-49, New, Noise



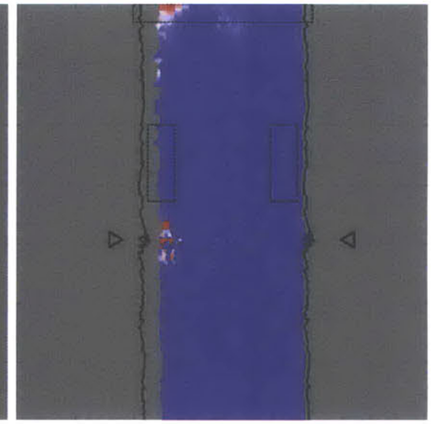
Scenario 08-13, New, Object



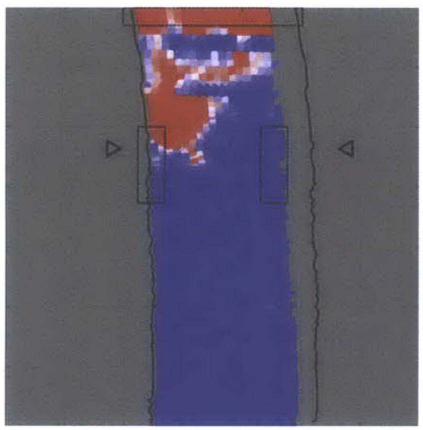
Scenario 08-16, New, Object



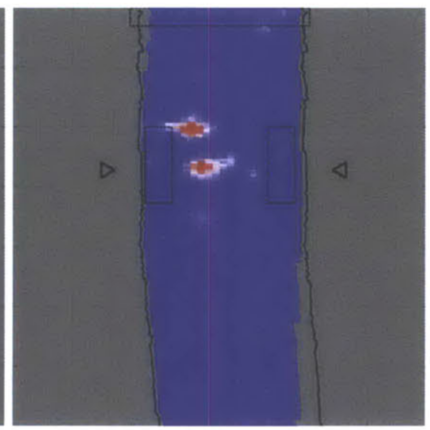
Scenario 08-17, New, Noise



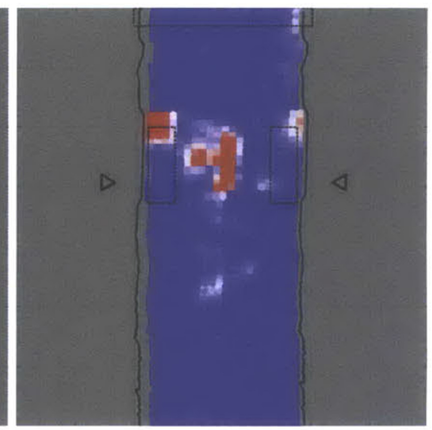
Scenario 08-37, New, Noise



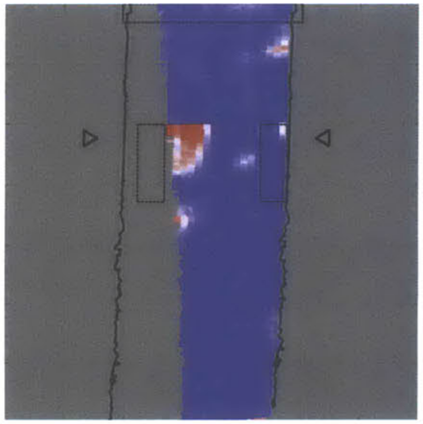
Scenario 09-15, New, Object



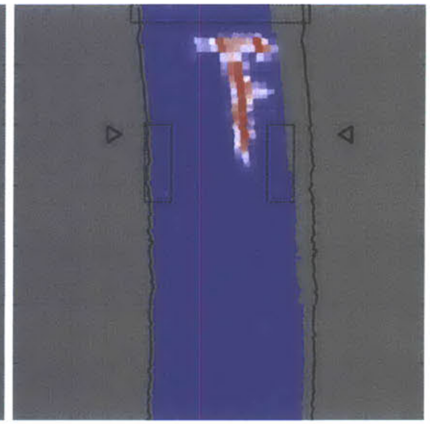
Scenario 09-17, New, Noise



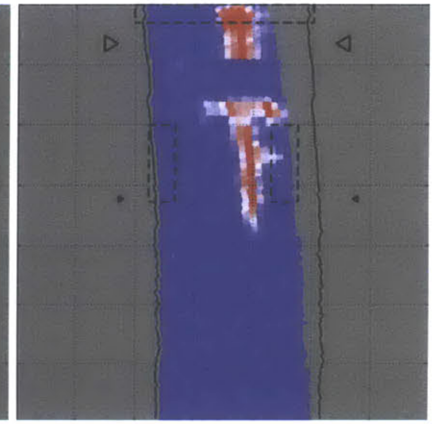
Scenario 09-18, New, Object



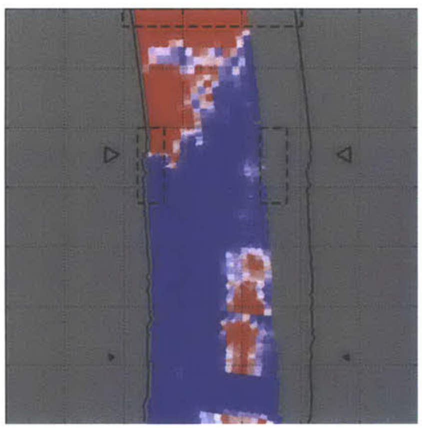
Scenario 09-39, New, Noise



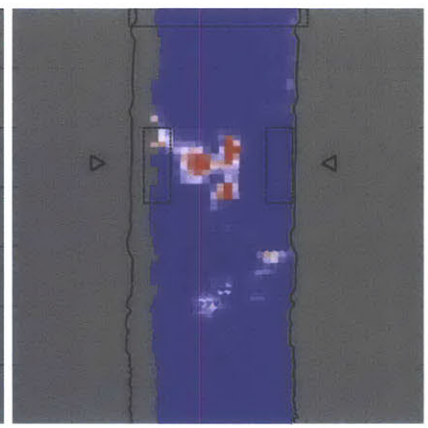
Scenario 10-15, New, Noise



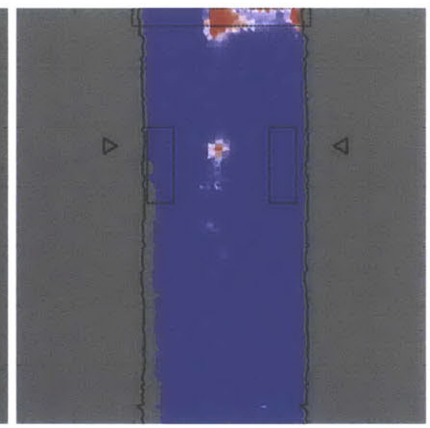
Scenario 10-16, New, Noise



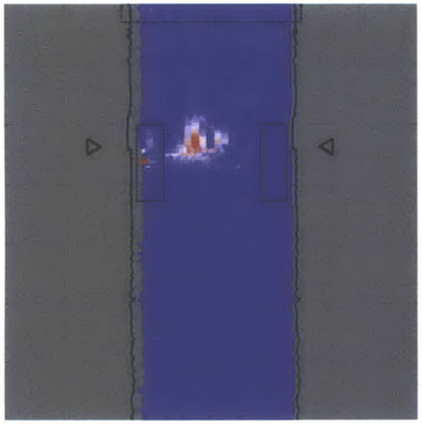
Scenario 10-17, New, Object



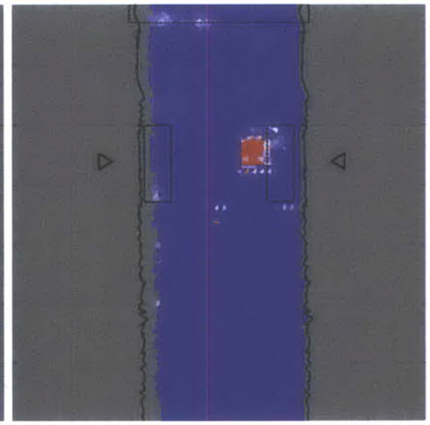
Scenario 10-21, New, Object



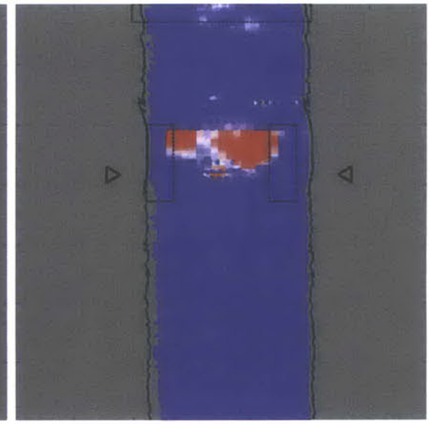
Scenario 10-30, New, Noise



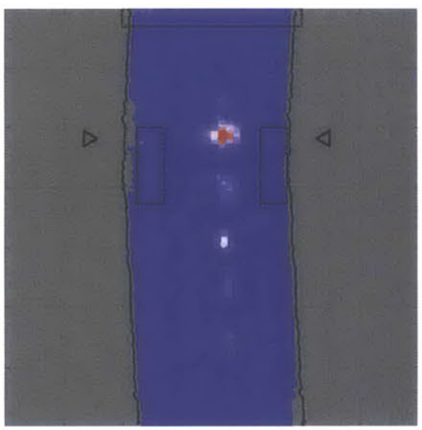
Scenario 10-34, New, Object



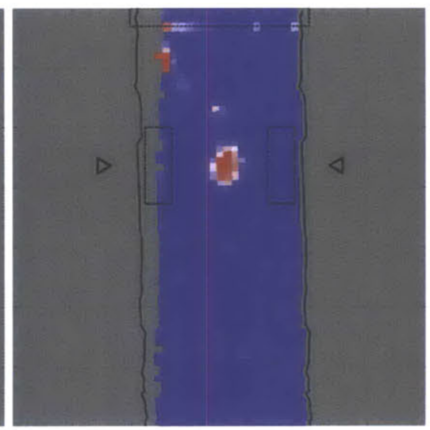
Scenario 10-37, New, Object



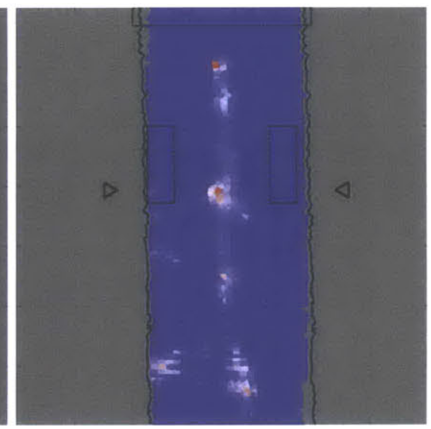
Scenario 10-38, New, Object



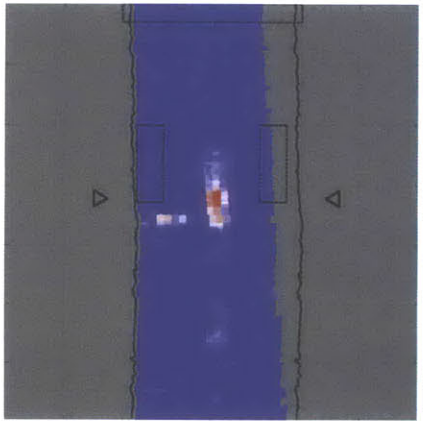
Scenario 11-02, New, Noise

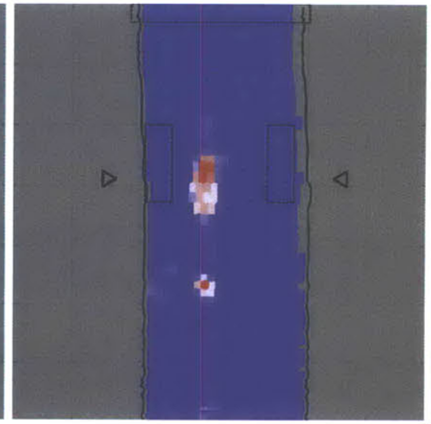


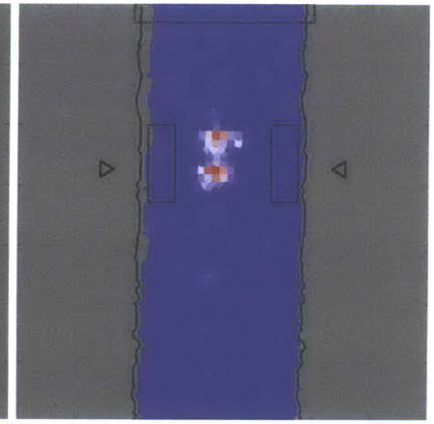
Scenario 11-05, New, Noise

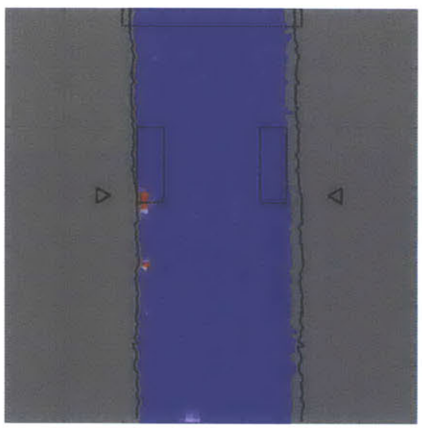


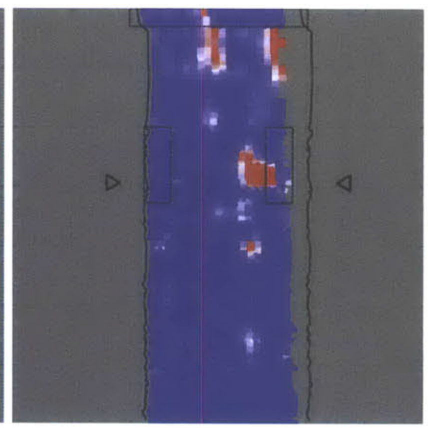
Scenario 11-36, New, Noise

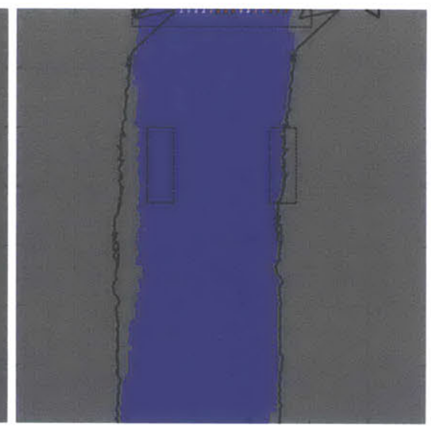

Scenario 11-38, New, Noise

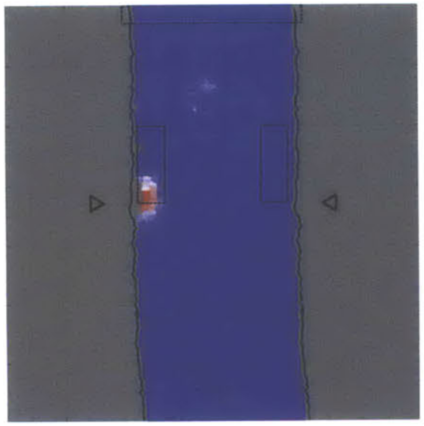

Scenario 12-15, New, Noise


Scenario 12-17, New, Noise

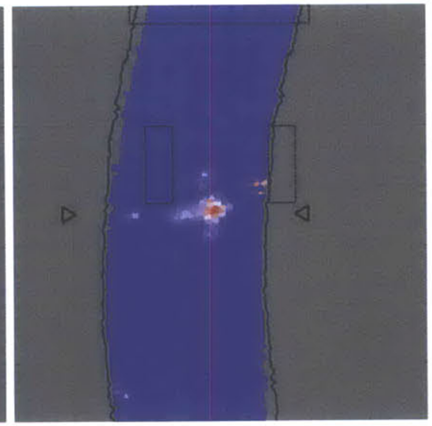

Scenario 12-25, New, Noise


Scenario 12-35, New, Object

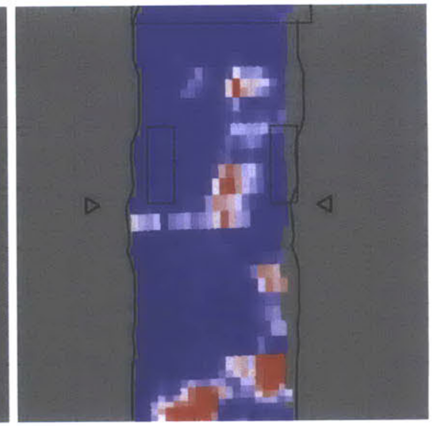

Scenario 12-40, New, Noise



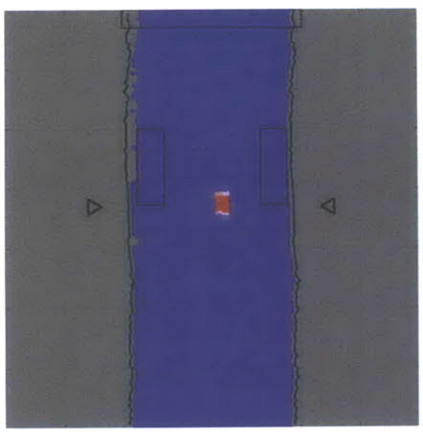
Scenario 13-18, New, Noise



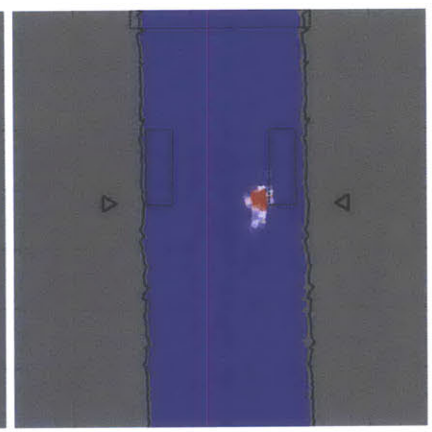
Scenario 13-22, New, Object



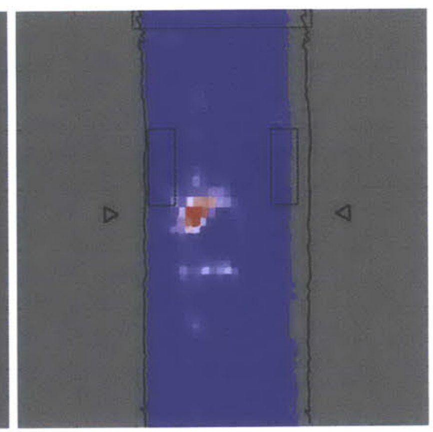
Scenario 14-15, New, Noise



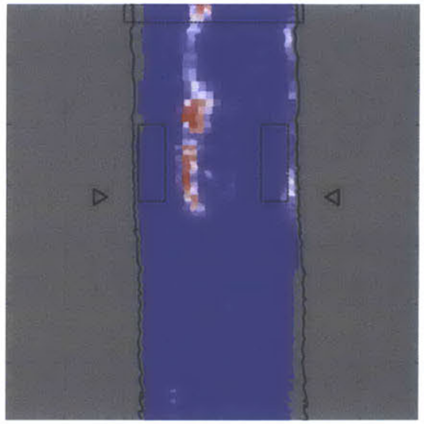
Scenario 14-16, New, Noise



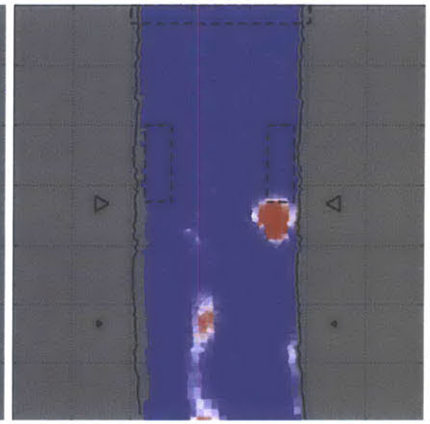
Scenario 14-17, New, Object



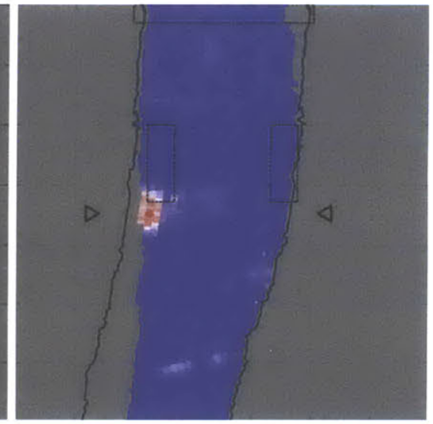
Scenario 14-24, New, Noise



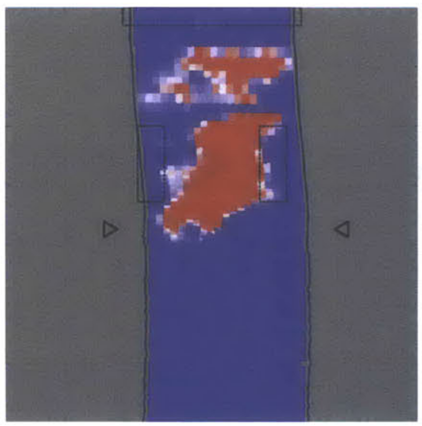
Scenario 14-25, New, Noise



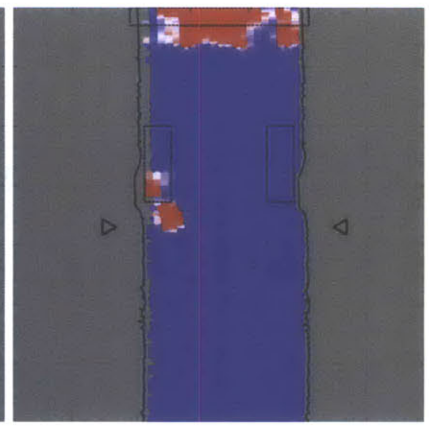
Scenario 14-27, New, Object



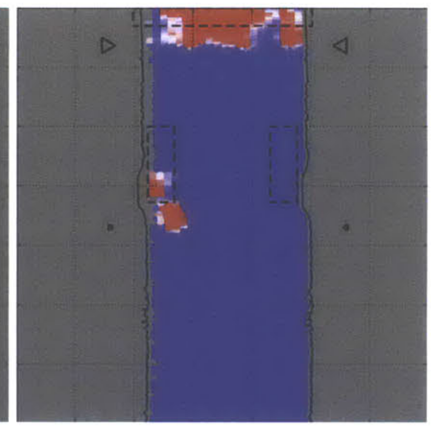
Scenario 14-36, New, Noise



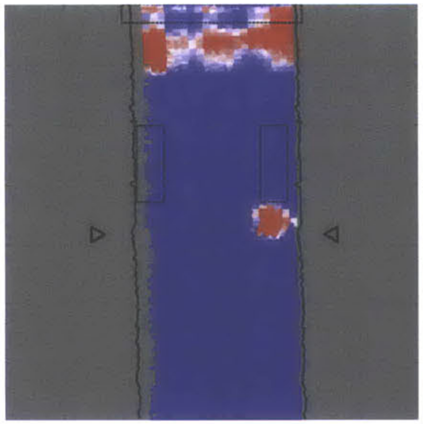
Scenario 15-07, New, Object



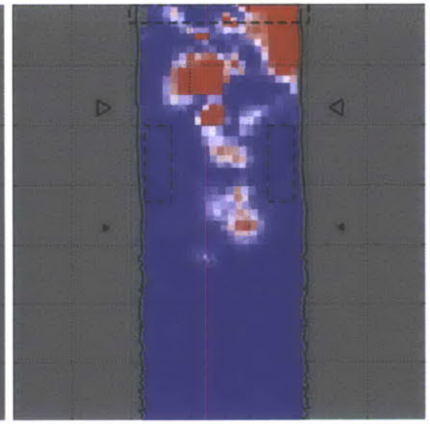
Scenario 16-04, New, Noise



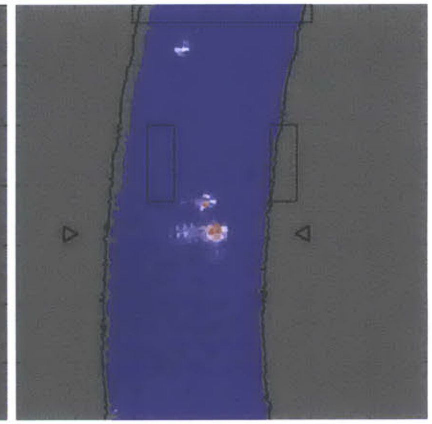
Scenario 16-05, New, Noise



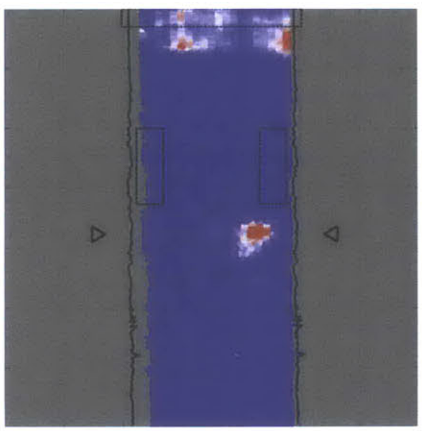
Scenario 16-15, New, Object



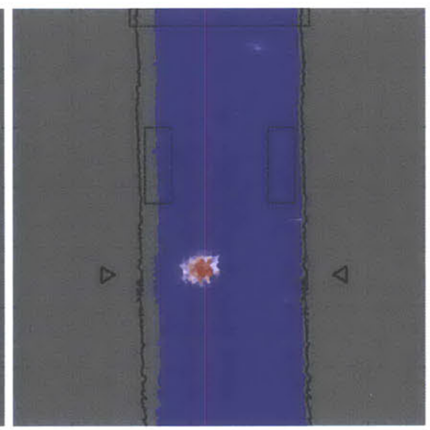
Scenario 16-20, New, Object



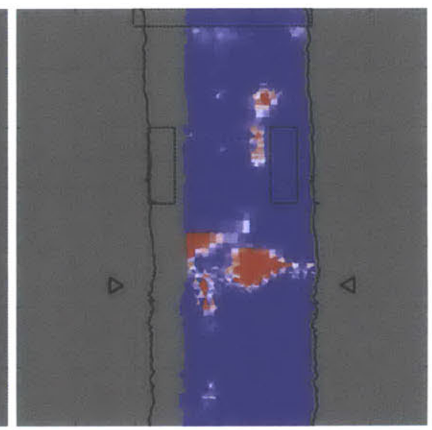
Scenario 16-28, New, Object



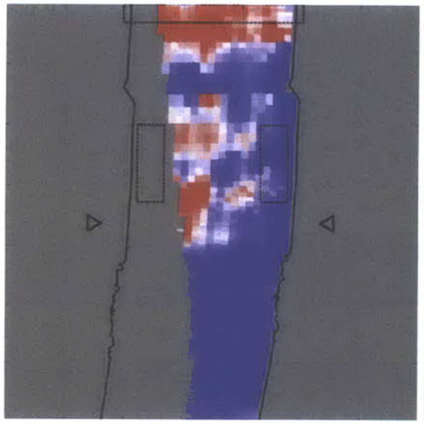
Scenario 16-30, New, Object



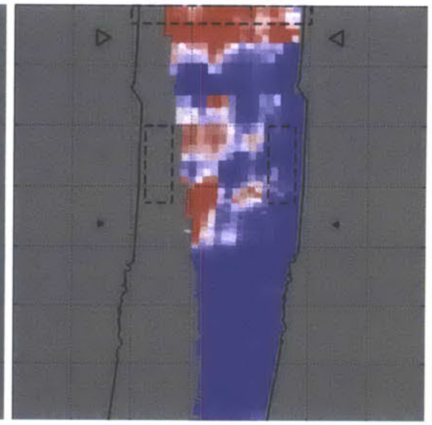
Scenario 16-37, New, Object



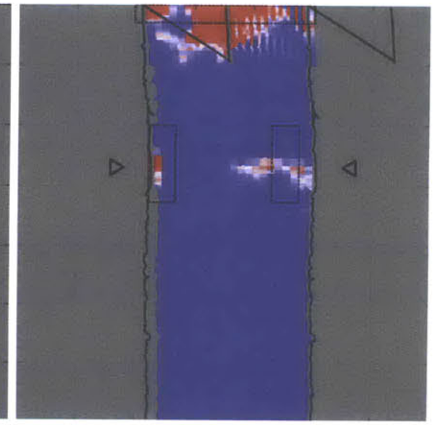
Scenario 16-42, New, Object



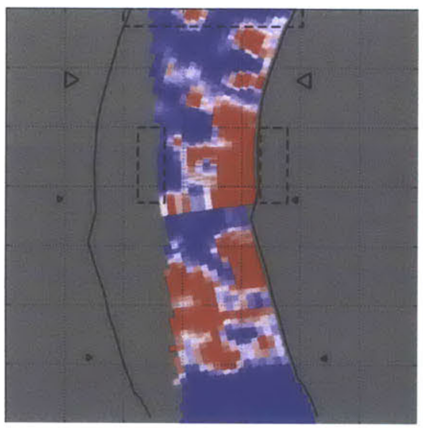
Scenario 16-43, New, Noise



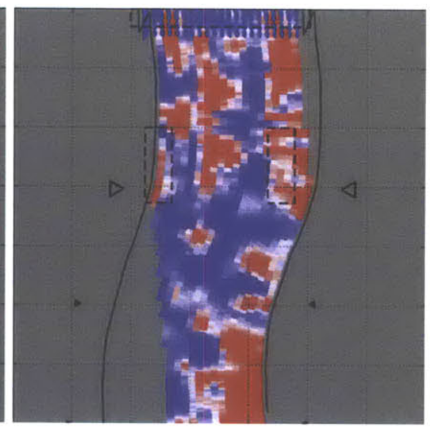
Scenario 16-44, New, Noise



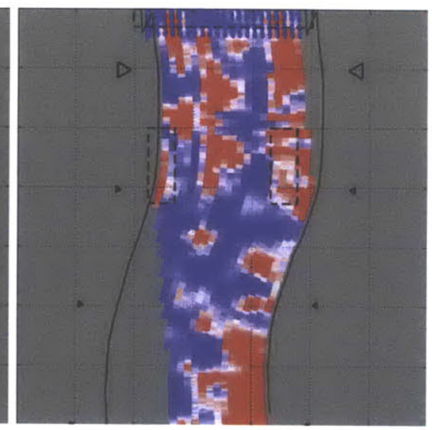
Scenario 17-08, New, Noise



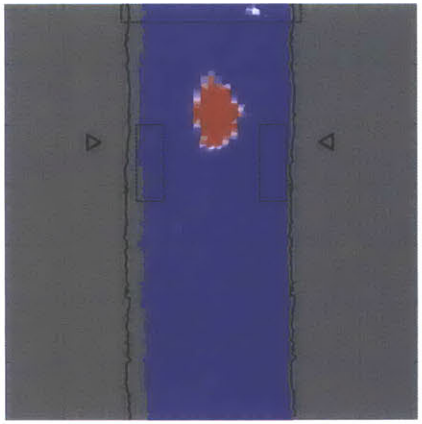
Scenario 17-12, New, Noise



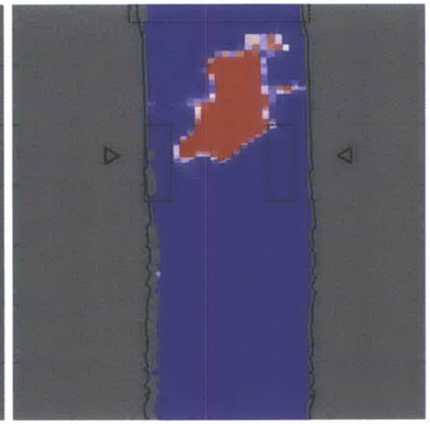
Scenario 17-13, New, Noise



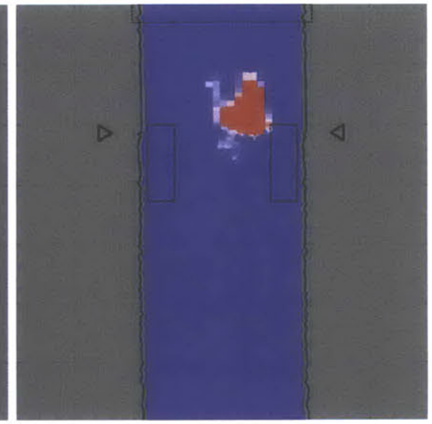
Scenario 17-14, New, Noise



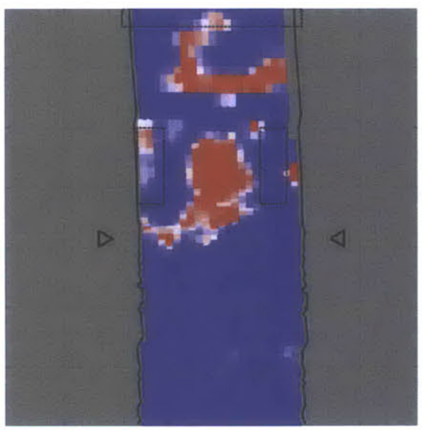
Scenario 18-05, New, Object



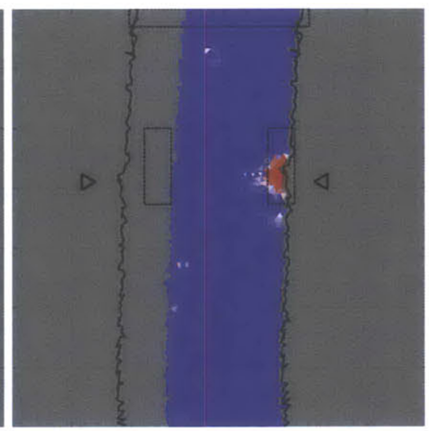
Scenario 18-07, New, Object



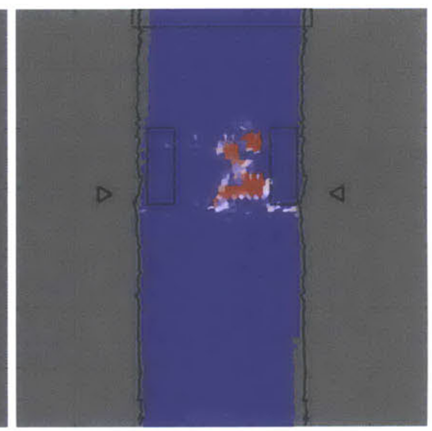
Scenario 18-09, New, Object



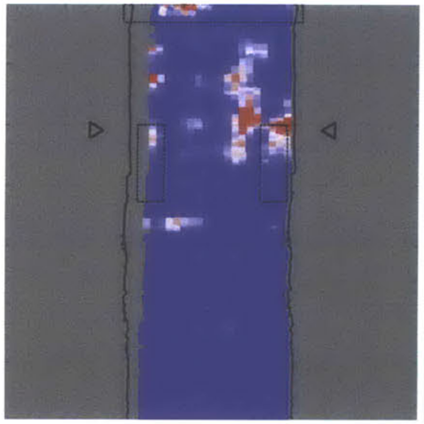
Scenario 19-07, New, Object



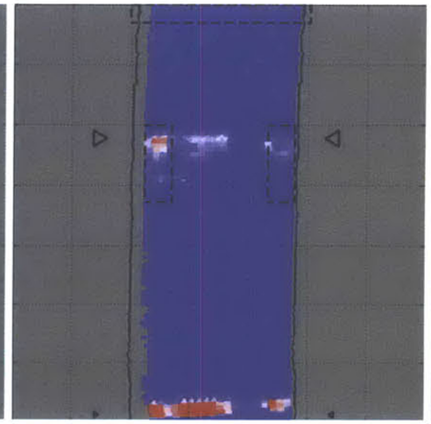
Scenario 21-02, New, Object



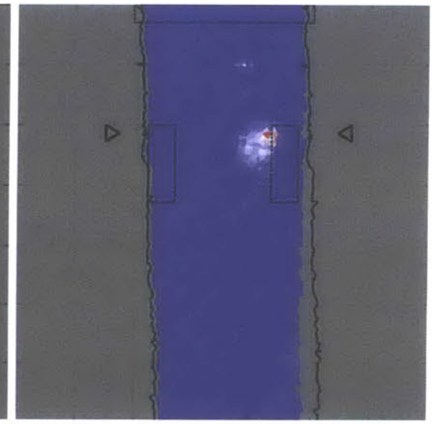
Scenario 21-03, New, Object



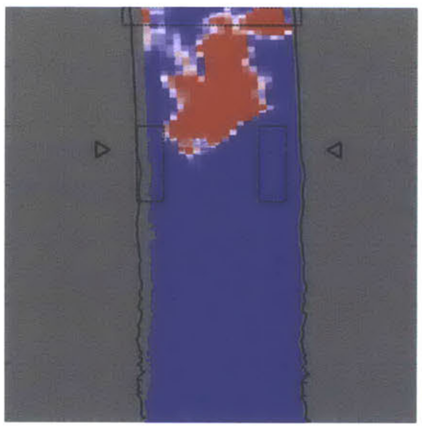
Scenario 21-06, New, Noise



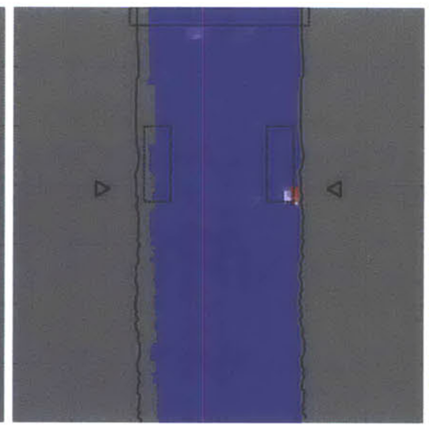
Scenario 21-09, New, Noise



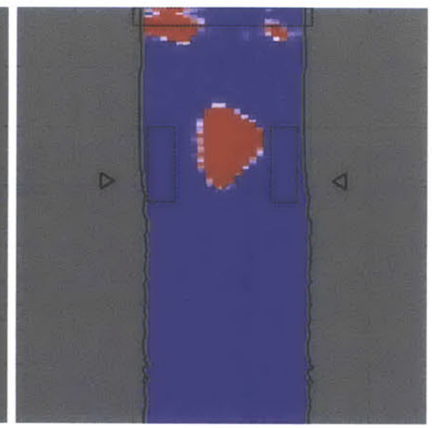
Scenario 21-10, New, Noise



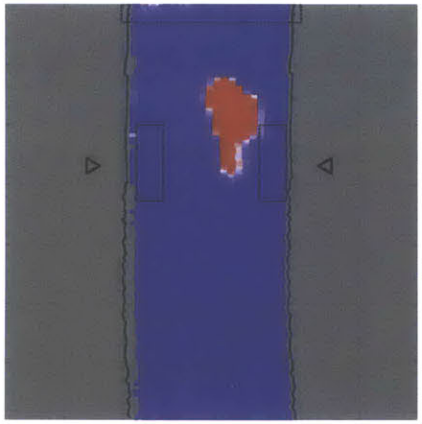
Scenario 21-11, New, Object



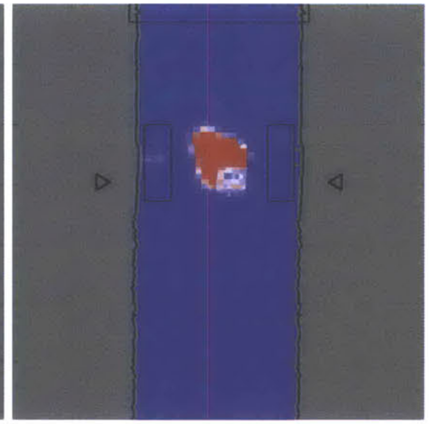
Scenario 21-14, New, Noise



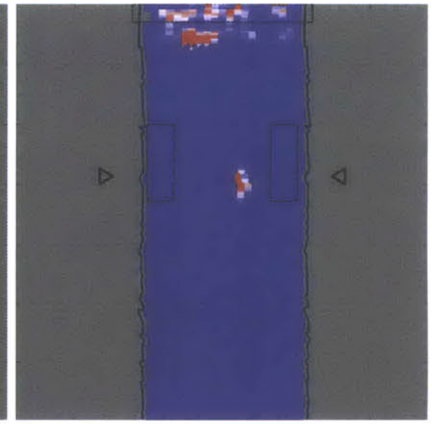
Scenario 21-16, New, Object



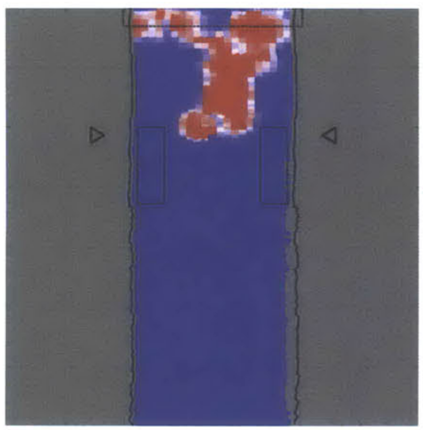
Scenario 21-18, New, Object



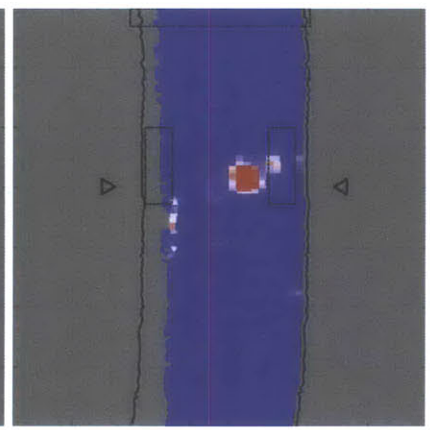
Scenario 22-03, New, Object



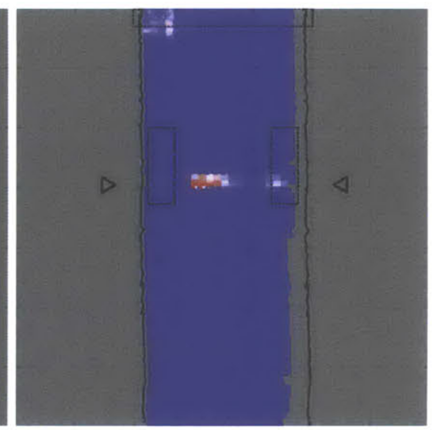
Scenario 22-04, New, Object



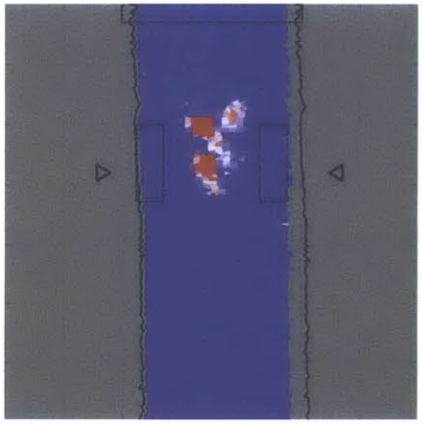
Scenario 22-10, New, Object



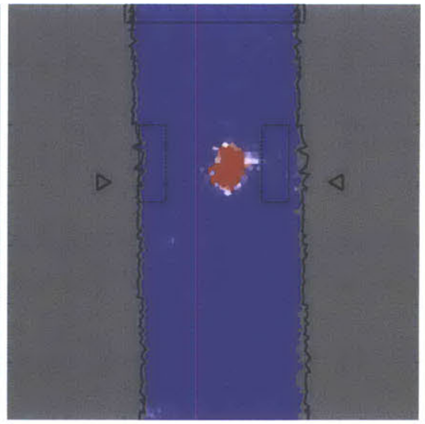
Scenario 22-14, New, Object



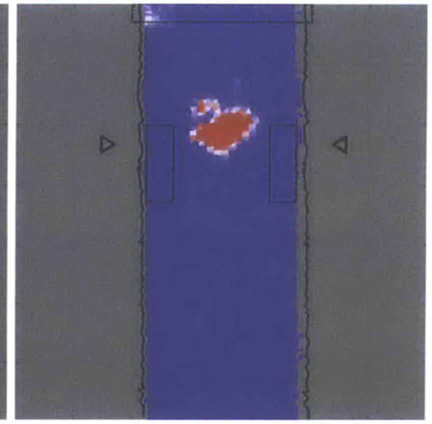
Scenario 22-16, New, Noise



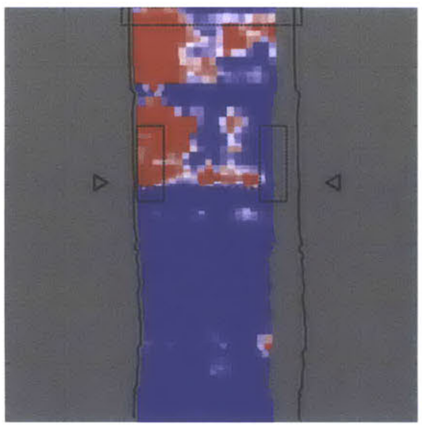
Scenario 22-20, New, Object



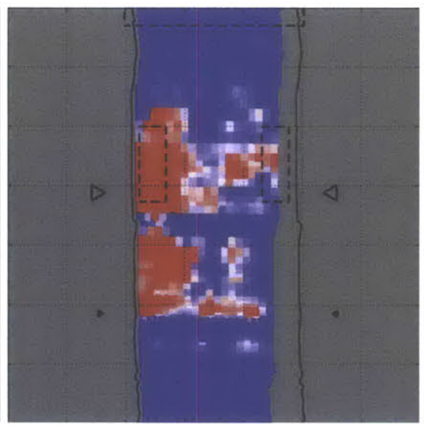
Scenario 22-21, New, Object



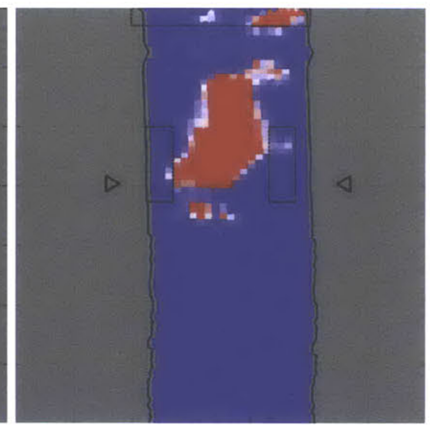
Scenario 22-22, New, Object



Scenario 25-07, New, Noise



Scenario 25-08, New, Noise



Scenario 25-09, New, Object

Appendix D

Data

Table D.1 and Table D.2 contain the raw data captured from subject surveys and parsed log files. In addition, a subset of the metrics analyzed in Chapter 5 have been included for convenience. All other metrics can be calculated from the data shown.

Subject #	Pilot Subject	First Interface	Expert	Job (0:Other; 1:Tech; 2:Mil; 3:Student)	Gender	Age (years)	Military XP	Colorblind	Touchscreen XP	Radar XP	Game XP (1: Least; 5: Most)	Conservativeness (1:Most; 5:Least)	Old: Est Correct %	New: Est Correct %	Interface Pref (1:Old; 5:New)	Old: TP	Old: TN	Old: FP	Old: FN	New: TP	New: TN	New: FP	New: FN
2	Y	1	N	1	M	22	N	N	4	N	2	2	67	75	4					33	32	11	15
4	Y	1	N	3	F	23	N	N	4	N	2	3		60	4					37	31	13	10
6	Y	2	N	1	M	24	N	N	4	Y	3	2	40	70	5					35	28	18	10
7	N	1	N	3	M	22	Y	N	4	Y	1	3	70	70	4	28	38	11	24	38	30	19	14
9	N	1	N	1	M	38	N	N	4	N	1	3	75	70	5	29	40	9	23	41	31	17	12
11	N	1	N	1	M	44	N	N	4	Y	1	2	80	60	4	34	35	14	18	42	27	22	10
13	N	1	N	1	M	68	N	N	4	Y	1	3	70	70	4	42	29	19	11	43	28	21	9
15	N	1	N	1	M	54	N	N	4	N	1	2	60	65	5	38	34	15	14	44	33	15	9
17	N	2	N	1	M	46	N	N	4	Y	1	2	50	60	5	37	36	13	15	32	33	16	20
19	N	2	N	1	M	29	N	N	3	Y	1	4	60	70	4	31	45	4	21	28	42	7	24
21	N	2	N	0	F	34	N	N	4	N	2	2	65	60	4	37	32	17	15	48	24	24	5
23	N	2	N	1	M	41	N	N	4	N	1	2	50	65	5	41	27	21	12	42	25	23	11
25	N	2	N	1	M	72	N	N	2	Y	1	3	60	75	3	22	43	5	31	24	36	12	29
27	N	2	N	1	M	35	N	N	4	Y	1	2	35	40	5	41	28	20	12	46	27	21	7
29	N	1	Y	2	M	48	Y	Y	4	N	3	4	96	50	4	20	43	6	33	44	23	26	9
31	N	2	Y	2	M	32	Y	N	4	Y	1	3	95	98	5	33	37	12	20	39	38	11	14
33	N	1	Y	1	M	32	N	N	4	Y	1	2	75	80	5	36	36	13	17	38	37	12	15
35	N	1	N	1	M	54	N	N	4	Y	1	2		80	5	33	34	15	20	34	36	13	19
37	N	1	N	0	M	42	N	N	3	N	5	2	70	50	1	39	32	17	14	31	31	18	22
39	N	2	N	1	F	37	N	N	3	Y	1	2	65	75	5	28	48	1	25	36	40	9	17
41	N	1	N	1	F	35	N	N	4	N	5	2	85	90	4	37	28	21	16	49	28	21	4
43	N	2	N	1	M	30	Y	N	3	Y	1	2	50	60	3	25	38	11	28	29	38	11	24
45	N	1	N	1	M	32	N	N	4	Y	3	4	70	75	5	33	35	14	20	32	33	16	21
47	N	1	N	1	M	63	N	N	4	N	1	3	50	65	5	23	35	14	30	36	33	16	17
49	N	2	N	1	M	34	N	N	3	N	2	2	70	60	4	36	36	13	17	43	30	19	10
51	N	1	N	1	M	33	N	N	4	N	4	2	75	50	2	32	40	9	21	41	30	19	12
53	N	2	N	1	M	35	N	N	4	Y	3	2	65	70	5	39	30	19	14	37	38	11	16
55	N	1	N	1	F	34	N	N	4	N	1	3	60	75	5	32	37	12	21	40	33	16	13
57	N	2	N	1	F	38	N	N	4	N	3	3	60	70	5	36	30	19	17	51	21	28	2
59	N	1	N	3	M	22	N	N	4	Y	1	2	75	75	5	35	31	18	18	40	33	16	13
61	N	2	N	1	M	27	N	N	4	Y	1	2	55	60	5	35	29	20	18	48	24	25	5
63	N	1	N	1	F	24	N	N	4	N	5	3	70	75	5	37	34	15	16	36	40	9	17
65	N	2	N	0	F	47	N	N	1	N	1	2	78	72	2	41	21	28	12	43	38	11	10
67	N	1	N	0	F	51	N	N	4	N	1	2	25	65	5	34	28	21	19	46	33	16	7
69	N	2	N	0	F	57	N	N	3	N	1	2	75	70	4	41	23	26	12	47	21	28	6
71	N	1	N	0	M	60	N	Y	4	N	1	4	50	75	5	41	32	17	12	38	27	22	15
73	N	2	N	0	M	54	N	N	1	N	1	3	67	72	5	23	37	12	30	38	24	25	15

Table D.1: Subject Survey Data and Metrics: Part 1

Subject #	Old: Correct %	New: Correct %	First: Correct %	Second: Correct %	Old: Mean Time Per Scenario (s)	New: Mean Time Per Scenario (s)	First: Mean Time Per Scenario (s)	Second: Mean Time Per Scenario (s)	Old: confidence-correctness	New: confidence-correctness	First: confidence-correctness	Second: confidence-correctness	Old: Average Confidence	New: Average Confidence	First: Average Confidence	Second: Average Confidence
2		71.4		71.4		3.85		3.85		0.64		0.64		4.81		4.81
4		74.7		74.7		0.85		0.85		0.49		0.49		4.00		4.00
6		69.2	69.2			5.34	5.34			0.46	0.46			5.27	5.27	
7	65.3	67.3	65.3	67.3	5.30	4.20	5.30	4.20	0.32	0.48	0.32	0.48	4.28	4.17	4.28	4.17
9	68.3	71.3	68.3	71.3	3.57	3.82	3.57	3.82	0.47	0.46	0.47	0.46	4.96	5.83	4.96	5.83
11	68.3	68.3	68.3	68.3	2.59	2.77	2.59	2.77	0.38	0.40	0.38	0.40	5.67	5.57	5.67	5.57
13	70.3	70.3	70.3	70.3	4.36	4.50	4.36	4.50	0.44	0.41	0.44	0.41	4.65	4.79	4.65	4.79
15	71.3	76.2	71.3	76.2	4.80	4.99	4.80	4.99	0.45	0.67	0.45	0.67	4.29	4.00	4.29	4.00
17	72.3	64.4	64.4	72.3	4.61	10.28	10.28	4.61	0.51	0.31	0.31	0.51	4.67	4.60	4.60	4.67
19	75.2	69.3	69.3	75.2	4.42	5.56	5.56	4.42	0.58	0.51	0.51	0.58	3.99	4.44	4.44	3.99
21	68.3	71.3	71.3	68.3	3.35	3.63	3.63	3.35	0.40	0.47	0.47	0.40	4.82	5.00	5.00	4.82
23	67.3	66.3	66.3	67.3	5.11	5.97	5.97	5.11	0.37	0.43	0.43	0.37	4.25	4.10	4.10	4.25
25	64.4	59.4	59.4	64.4	3.11	3.49	3.49	3.11	0.33	0.30	0.30	0.33	4.25	4.18	4.18	4.25
27	68.3	72.3	72.3	68.3	3.61	5.05	5.05	3.61	0.39	0.46	0.46	0.39	3.34	3.87	3.87	3.34
29	61.8	65.7	61.8	65.7	5.19	9.00	5.19	9.00	0.27	0.37	0.27	0.37	6.34	6.05	6.34	6.05
31	68.6	75.5	75.5	68.6	4.16	4.98	4.98	4.16	0.40	0.55	0.55	0.40	5.14	5.13	5.13	5.14
33	70.6	73.5	70.6	73.5	3.86	3.46	3.86	3.46	0.49	0.54	0.49	0.54	4.65	4.46	4.65	4.46
35	65.7	68.6	65.7	68.6	14.92	10.57	14.92	10.57	0.36	0.46	0.36	0.46	4.19	4.04	4.19	4.04
37	69.6	60.8	69.6	60.8	3.08	4.12	3.08	4.12	0.40	0.30	0.40	0.30	3.68	2.93	3.68	2.93
39	74.5	74.5	74.5	74.5	6.85	8.01	8.01	6.85	0.48	0.55	0.55	0.48	4.62	4.81	4.81	4.62
41	63.7	75.5	63.7	75.5	3.40	3.19	3.40	3.19	0.37	0.60	0.37	0.60	3.87	4.25	3.87	4.25
43	61.8	65.7	65.7	61.8	3.57	4.27	4.27	3.57	0.32	0.40	0.40	0.32	4.27	4.58	4.58	4.27
45	66.7	63.7	66.7	63.7	5.53	5.10	5.53	5.10	0.41	0.36	0.41	0.36	4.59	4.59	4.59	4.59
47	56.9	67.6	56.9	67.6	4.25	3.61	4.25	3.61	0.13	0.36	0.13	0.36	4.98	5.03	4.98	5.03
49	70.6	71.6	71.6	70.6	2.10	4.11	4.11	2.10	0.43	0.47	0.47	0.43	4.06	4.33	4.33	4.06
51	70.6	69.6	70.6	69.6	2.94	2.69	2.94	2.69	0.48	0.55	0.48	0.55	4.02	3.78	4.02	3.78
53	67.6	73.5	73.5	67.6	3.24	3.28	3.28	3.24	0.42	0.54	0.54	0.42	4.13	4.50	4.50	4.13
55	67.6	71.6	67.6	71.6	3.55	2.39	3.55	2.39	0.41	0.55	0.41	0.55	4.33	5.18	4.33	5.18
57	64.7	70.6	70.6	64.7	3.13	4.03	4.03	3.13	0.35	0.45	0.45	0.35	5.46	5.82	5.82	5.46
59	64.7	71.6	64.7	71.6	3.30	2.29	3.30	2.29	0.39	0.53	0.39	0.53	4.92	4.62	4.92	4.62
61	62.7	70.6	70.6	62.7	3.12	3.83	3.83	3.12	0.37	0.49	0.49	0.37	4.42	4.81	4.81	4.42
63	69.6	74.5	69.6	74.5	3.40	2.11	3.40	2.11	0.40	0.57	0.40	0.57	5.65	5.96	5.65	5.96
65	60.8	79.4	79.4	60.8	2.95	4.08	4.08	2.95	0.22	0.58	0.58	0.22	4.72	4.66	4.66	4.72
67	60.8	77.5	60.8	77.5	8.77	5.79	8.77	5.79	0.23	0.62	0.23	0.62	4.15	4.52	4.15	4.52
69	62.7	66.7	66.7	62.7	3.00	3.09	3.09	3.00	0.31	0.37	0.37	0.31	5.46	5.93	5.93	5.46
71	71.6	63.7	71.6	63.7	3.91	3.45	3.91	3.45	0.44	0.33	0.44	0.33	4.83	4.55	4.83	4.55
73	58.8	60.8	60.8	58.8	6.73	12.44	12.44	6.73	0.20	0.22	0.22	0.20	5.34	4.92	4.92	5.34

Table D.2: Subject Survey Data and Metrics: Part 2

	Old: Est Correct %	New: Est Correct %	Interface Pref (1:Old; 5:New)	Old: TP	Old: TN	Old: FP	Old: FN	New: TP	New: TN	New: FP	New: FN	Old: Correct %	New: Correct %	First: Correct %	Second: Correct %
Mean	64.66	68.16	4.32	33.79	34.15	14.74	19.00	39.16	31.24	17.22	13.19	66.83	69.85	68.13	68.68
Median	66	70	5	35	34	15	18	39	31	17	13	67.65	70.44	68.77	68.63
Mode	70	75	5	41	35	14	12	38	33	16	10				
Min	25	40	1	20	21	1	11	24	21	7	2	56.86	59.41	56.86	58.82
Max	96	98	5	42	48	28	33	51	42	28	29	75.25	79.41	79.41	77.45
Std	15.13	11.01	1.00	6.04	5.97	5.93	6.04	6.27	5.57	5.64	6.13	4.39	4.78	4.77	4.90

	Old: Mean Time Per Scenario (s)	New: Mean Time Per Scenario (s)	First: Mean Time Per Scenario (s)	Second: Mean Time Per Scenario (s)	Old: confidence-correctness	New: confidence-correctness	First: confidence-correctness	Second: confidence-correctness	Old: Average Confidence	New: Average Confidence	First: Average Confidence	Second: Average Confidence
Mean	4.40	4.71	5.09	4.05	0.38	0.47	0.41	0.44	4.62	4.71	4.71	4.61
Median	3.61	4.10	4.18	3.61	0.39	0.47	0.42	0.42	4.62	4.61	4.65	4.59
Mode												
Min	2.10	0.85	2.59	0.85	0.13	0.22	0.13	0.20	3.34	2.93	3.68	2.93
Max	14.92	12.44	14.92	10.57	0.58	0.67	0.58	0.67	6.34	6.05	6.34	6.05
Std	2.30	2.46	2.72	1.88	0.09	0.11	0.10	0.12	0.64	0.68	0.62	0.70

Table D.3: Metric Summary Statistics

Appendix E

Algorithms

E.1 GPR Rendering

The GPR displays for both interfaces were rendered in Matlab®. They were rendered using data files which provided GPR data in addition to the location and position of the radar array. The hardware for the interface allowed an 800x480 pixel display, which was chosen as the size for rendering on the iPad®implementation. See Figure 3-7 for the colormaps used in both display.

In the case of the old display, the data was displayed using rectangles with each rectangle representing one sensor value as read from the data files. No data preprocessing was needed. The colormap used a 65-value color system where the lowest value was indicative of a lack of data. The remaining 64 values were drawn from a gradient colormap.

The in the case of the new display, the data were displayed using individual pixels at native resolution for the iPad®. For each detection on this display the data were preprocessed and rendered as follows:

1. The stopping location and time of the vehicle were identified.
2. All of the GPR sensor values were transformed into a cartesian coordinate system.
3. The individual locations of every sensor reading visible in the end product was calculated using the geometry of the sensors and vehicle.
4. The data were rotated around the stopping location such that the heading of the vehicle pointed vertically.
5. A nearest-neighbor scattered interpolation model was calculated from the data.
6. The display was rendered using the nearest-neighbor interpolation at the native resolution of the iPad® screen such that each interpolated point corresponded to one pixel on the final image.
7. The graphical elements of the interface discussed in Section 3.5 were rendered on top of the GPR data.

E.2 Braking

Gathering a large data set for this experiment was not easy. The data used in this experiment was captured during hardware testing for other purposes over a period of several days. One consequence of this method of collection was that in some of the scenarios, the automatic braking system had been deactivated. In order to include more examples in this study, detection data captured from the non-braking data set was processed and used in addition to the data captured when the system did brake.

In scenarios where the system did brake, the displays (found in Appendix C) were rendered using data captured up to the time when the system came to a halt.

In scenarios where the system did not brake automatically, a linear model of braking distance as a function of speed was used to estimate the position where the system would have ended up if it had braked. The braking function was determined by examining the examples in which the system did brake and finding the finding a linear regression between the system speed at the start of braking and the distance it took to stop. Then that regression was used to predict braking distance for the scenarios that did not involve braking. There were no observable differences between the renderings of real and simulated braking scenarios. 58 of the 138 potential detections in the data set were simulated.

E.3 Boosting

E.3.1 Artificial Intelligence

If two different classifiers accept the same input, they can be combined into a third classifier in a processing known as boosting. The goal of boosting is usually to create a 'strong' classifier out of multiple 'weak' classifiers. There are many boosting algorithms that use slightly different mathematical approaches to combine classifiers. Two well-known boosting algorithms are AdaBoost and Eta-Boost.

It is not always that case that two or more classifiers can be combined in a useful manner. Consider the following example scenarios involving two classifiers, A and B shown in Figure E-1.

Classifier A and B each correctly classify the same number of the inputs, 8 of 12. In Scenario 1, A and B demonstrate substantial non-overlapping expertise; they get different inputs correct. In Scenario 2, A and B demonstrate overlapping expertise; they get the same inputs correct. Consider the result of boosting A and B in these cases. In Scenario 1, each classifier contributes some inputs that it correctly classifies which are novel to the hybrid. If it were possible to predict the likelihood that A and B would correctly classify an input, then a hybrid classifier could use the following algorithm:

Algorithm 1 attempts to select the single best classifier for a particular input.

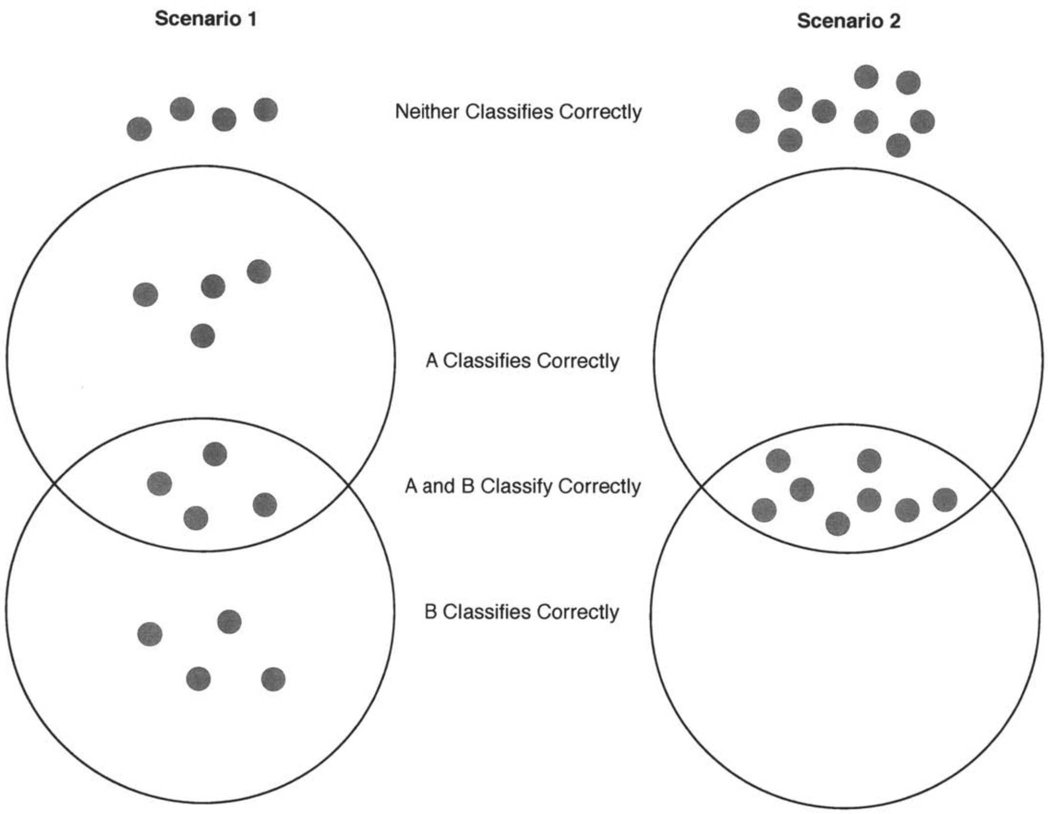


Figure E-1: Boosting Examples

Other strategies involve combining several classifiers.

Boosting transforms the difficulty of classifying an input into two subproblems: classifying inputs by several different means and determining which classifiers are best suited for a given input. If these two problems are more tractable than the original problem, then boosting can be a valuable tool.

E.3.2 Human Subjects Boosting

Boosting for human classifiers is the same theoretical problem as boosting for artificial intelligence classifiers. Rather than picking the single best subject as in Algorithm 1, Algorithm 2 was used.

In contrast to Algorithm 1, this algorithm assumes that all classifiers are equally

Algorithm 1 Boosting a collection of weak classifiers by choosing the best

```
function EvaluateClassifierFitness(classifier, input)
    return A number indicating the goodness of this classifier by
           previously observed performance in similar situations
end function

function SingleBestBoostedClassify(classifiers, input)
    fitness ← [ ]
    for all c ∈ classifiers do
        fitness[c] ← EvaluateClassifierFitness(c,input)
    end for
    best_classifier ← argmax(fitness)
    return best_classifier(input)
end function
```

Algorithm 2 Boosting human classifiers by committee voting

```
function CommitteeBoostedClassify(subjects, scenario, threshold)
    stimulus ← 0
    for all s ∈ subjects do
        stimulus ← stimulus + SignExtendedSubjectConfidence(s,scenario)
    end for
    if stimulus ≥ threshold then
        return True
    else
        return False
```

fit. Everyone votes on the outcome. Algorithm 2 uses the sign-extended confidence scores of each subject. Sign-extended confidence scores were calculated by taking the confidence response from each subject $[1 - 7]$ and multiplying by $+1$ if the subject believed the detection was an object or multiplying by -1 if the subject believed the detection was a false alarm, yielding an integer score $[-7, 7]$ excluding 0. Summing these sign-extended confidence scores over all the committee members provides an integer vote tally. If the tally was above the chosen decision threshold, then the committee decides the signal was an object. If the tally was below the threshold, then the committee decides the signal was a false alarm.

There are many potential variations on this algorithm. One concern with simply adding up confidences is that each subject may have a different distribution of overall confidence. For instance, some subjects might have more extreme confidences while others responded more centrally. A potential solution to this would be to run each subject's sign extended confidence score through a personalized function to normalize that subject's scores. Trivial normalization algorithms, such as dividing by each subject's average original confidence score, did not demonstrate significant improvement.

References

[1] A. P. Annan, "GPR---History, Trends, and Future Developments," *Subsurface Sensing Technologies and Applications*, vol. 3, no. 4, pp. 253--270, 2002.

[2] H. Herman, J. D. McMahill, and G. Kantor, "Enhanced operator interface for hand-held landmine detector," *Proceedings of the Society for Optics and Photonics*, vol. 4394, pp. 844--851, 2001.

[3] "Cause and Effect: The Grid," *Underground Focus Magazine*, vol. 26, no. 4, 2012.

[4] UPI, "THOUSANDS EVACUATED AFTER COAST GAS LINE BURSTS," *New York Times*, 1981.

[5] A. Gardner, "Near Washington, D.C., construction crews watch for mystery 'black' wire," *Los Angeles Times*, 2009.

[6] A. O. Nasif, K. J. Hintz, and N. Peixoto, "Syntactic landmine detection and classification," *Proceedings of the Society for Optics and Photonics*, vol. 7664, pp. 76642F--76642F--10, 2010.

[7] A. J. Bosker, "IEDs will remain 'weapon of choice' for decades," *Joint IED Defeat Organization News Service*, 2012.

[8] J. Francke, "Applications of GPR in mineral resource evaluations," in *Proceedings of the XIII Internarional Conference on Ground Penetrating Radar*, pp. 1--5, IEEE, 2010.

[9] A. Catakli, H. Mahdi, and H. Al-Shukri, "Texture analysis of GPR data as a tool for depicting soil mineralogy," in *2011 IEEE Applied Imagery Pattern Recognition Workshop*, pp. 1--8, IEEE, 2011.

[10] A. Neal and C. L. Roberts, "Applications of ground-penetrating radar (GPR) to sedimentological, geomorphological and geoarchaeological studies in coastal environments," *Geological Society, London, Special Publications*, vol. 175, no. 1, pp. 139--171, 2000.

[11] D. Eisenburger, H. Lentz, and M. Jenett, "Helicopter-borne GPR systems: A way from ice thickness measurements to geological applications," in *2008 IEEE International Conference on Ultra-Wideband*, pp. 161--165, IEEE, 2008.

[12] J. E. Lucius, "Detectability Of Crude Oil In The Subsurface Near Bemid Ji, Minnesota, Using Ground Penetrating Radar," in *13th EEGS Symposium on the Application of Geophysics to Engineering and Environmental Problems*, 2000.

[13] B. L. Burton, G. R. Olhoeft, and M. H. Powers, "Frequency spectral analysis of GPR data over a crude oil spill," in *Proceedings of the X International Conference on Ground Penetrating Radar*, pp. 267--270, 2004.

[14] K. O'Neill, K. Sun, C. C. Chen, F. Shubitidze, and K. D. Paulsen, "Combining GPR and EMI data for discrimination of multiple subsurface metallic objects," in *2003 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4157--4159, IEEE, 2003.

[15] H. Herman and D. Iglesias, "Human-in-the-loop issues for demining," *Proceedings of the Society for Optics and Photonics*, vol. 3710, pp. 797--805, 1999.

[16] H. Herman, J. D. McMahill, and G. Kantor, "Training and performance assessment of land mine detector operator using motion tracking and virtual mine lane," *Proceedings of the Society for Optics and Photonics*, vol. 4038, pp. 110--121, 2000.

[17] H. Herman, T. Higgins, O. Falmier, J.-S. Valois, and J. McMahill, "Mine detection performance comparison between manual sweeping and tele-operated robotic system," *Proceedings of the Society for Optics and Photonics*, vol. 7664, pp. 766419--766419--12, 2010.

[18] J. N. Wilson, P. Gader, K. C. Ho, and R. Mazhar, "An analysis of sweep patterns for a handheld demining system," *Proceedings of the Society for Optics and Photonics*, vol. 6217, pp. 62172W--62172W--12, 2006.

[19] M. Reddy, S. Agarwal, R. Hall, J. Brown, T. Woodard, and A. Trang, "Warfighter-in-the-loop: mental models in airborne minefield detection," *Proceedings of the Society for Optics and Photonics*, vol. 5794, pp. 1050--1059, 2005.

[20] Y. Tan, S. L. Tantum, and L. M. Collins, "Enhanced signal and auditory processing for landmine detection using EMI sensors," *Proceedings of the Society for Optics and Photonics*, vol. 4394, pp. 852--858, 2001.

[21] S. Throckmorton, Y. Tan, P. Gao, L. Gresham, and L. M. Collins, "Enhanced auditory processing for land mine detection using EMI sensors," *Proceedings of the Society for Optics and Photonics*, vol. 3710, pp. 787--796, 1999.

[22] N. L. Vause, T. R. Letowski, L. G. Ferguson, and T. J. Mermagen, "Auditory issues in handheld land mine detectors," *Proceedings of the Society for Optics and Photonics*, vol. 3710, pp. 778--786, 1999.

[23] B. M. Davis, W. W. Winchester, and J. D. Zedlitz, "Auditory Visualization of Landmine Detector Sensor Data," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53, no. 18, pp. 1304--1308, 2009.

[24] Y. Tan, L. Huettel, and L. M. Collins, "Predicting improved human auditory discrimination for land mine detection using EMI sensors," *Proceedings of the Society for Optics and Photonics*, vol. 4038, pp. 122--129, 2000.

[25] J. J. Staszewski and A. Davison, "Mine detection training based on expert skill," *Proceedings of the Society for Optics and Photonics*, vol. 4038, pp. 90--101, 2000.

[26] J. J. Staszewski, "Models of Human Expertise as Blueprints for Cognitive Engineering: Applications to Landmine Detection," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 3, pp. 458--462, 2004.

[27] R. Drazovich, S. Brooks, and S. Foster, "Knowledge Based Ship Classification," tech. rep., Palo Alto, California, 1979.

[28] L. Varshney, "Human Machine Interface for Radar Systems," tech. rep., North Syracuse, New York, 2002.

[29] C. E. Nehme, S. D. Scott, M. L. Cummings, and C. Y. Furusho, "Generating Requirements for Futuristic Hetrogenous Unmanned Systems," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 3, pp. 235--239, 2006.

[30] L. D. Hunt, D. Massie, and J. P. Cull, "Standard palettes for GPR data analysis," *Proceedings of the Society for Optics and Photonics*, vol. 4084, pp. 341--345, 2000.

[31] M. Harrower and C. A. Brewer, "ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27--37, 2003.

[32] J. M. Rathje, L. B. Spence, and M. L. Cummings, "Human-Automation Collaboration in Occluded Trajectory Smoothing," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 137--148, 2013.

This work is sponsored by the
Assistant Secretary of Defense for Research & Engineering
under Air Force Contract #FA8721-05-C-0002.